NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

EVALUATION OF FACTORS ON THE PATTERNS OF SHIP MOVEMENT AND PREDICTABILITY OF FUTURE SHIP LOCATION IN THE GULF OF MEXICO

by

Sophia M. Bay

March 2017

Thesis Advisor: Robert A. Koyak
Second Reader: Jonathan K. Alt

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** March 2017	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE
EVALUATION OF FACTORS ON THE PATTERNS OF SHIP MOVEMENT AND PREDICTABILITY OF FUTURE SHIP LOCATION IN THE GULF OF MEXICO

5. FUNDING NUMBERS

6. AUTHOR(S) Sophia M. Bay

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB number ____N/A____.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release. Distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (maximum 200 words)

In this thesis, we examine techniques used to predict future ship movement using historical Automatic Identification System (AIS) data in the Gulf of Mexico from April 2014. We process the data to remove outliers and identify "subtracks," which are associated with trips made by a vessel between two points. A cluster analysis is then used to determine the extent to which subtrack routes segregate into groups in an area without well-defined shipping lanes. Although clustering structure does exist, it is not strong enough to support prediction modeling in line with other published work. We also examine the effects of weather and sea-state on deviations of a vessel's traveled route from the shortest (great-circle) route. Vessels vary substantially in how closely they adhere to a great-circle route. Head winds also contribute positively to these deviations. This result suggests that algorithms designed to predict the motion of vessels should take weather and sea-state into account.

14. SUBJECT TERMS
data analysis, Automatic Information System, Gulf of Mexico shipping traffic

15. NUMBER OF PAGES
81

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

EVALUATION OF FACTORS ON THE PATTERNS OF SHIP MOVEMENT AND PREDICTABILITY OF FUTURE SHIP LOCATION IN THE GULF OF MEXICO

Sophia M. Bay
Lieutenant, United States Navy
B.S., The Citadel, 2010

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
March 2017

Approved by: Robert A. Koyak
Thesis Advisor

Jonathan K. Alt
Second Reader

Patricia Jacobs
Chair, Department of Operations Analysis

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

In this thesis, we examine techniques used to predict future ship movement using historical Automatic Identification System (AIS) data in the Gulf of Mexico from April 2014. We process the data to remove outliers and identify “subtracks,” which are associated with trips made by a vessel between two points. A cluster analysis is then used to determine the extent to which subtrack routes segregate into groups in an area without well-defined shipping lanes. Although clustering structure does exist, it is not strong enough to support prediction modeling in line with other published work. We also examine the effects of weather and sea-state on deviations of a vessel’s traveled route from the shortest (great-circle) route. Vessels vary substantially in how closely they adhere to a great-circle route. Head winds also contribute positively to these deviations. This result suggests that algorithms designed to predict the motion of vessels should take weather and sea-state into account.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AIC	Akaike Information Criterion
AIS	Automatic Identification System
BIC	Bayes Information Criterion
BOEM	Bureau of Ocean Energy Management
CROSS	crosswind component of wind speed in meters/second, averaged over the subtracks
DOWN	downwind component of wind speed in meters/second, averaged over the subtracks
HSC	high speed vessel
IMO	International Maritime Organization
LOOP	Louisiana Offshore Oil Platform
MDA	maritime domain awareness
MMSI	Maritime Mobile Service Identity
NPS	Naval Postgraduate School
WIG	wing in ground
WLS	weighted least squares
WVHT	wave height in meters

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

Over the past few decades the number of sea-going vessels has increased substantially which poses a challenge to maintain situational awareness of the global maritime picture. In 2002 the International Maritime Organization (IMO) implemented the Automatic Identification System (AIS) which allows a vessel to broadcast its position, movement, and static information about the vessel. Since its debut, AIS uses have expanded to include the monitoring of fishing vessels, search and rescue, meteorological data, and maritime security. Information provided by AIS includes speed, heading, latitude and longitude, ship type, ship dimensions, and destination as well as other attributes. Although voluminous, AIS data is archived and made available to the public. Because of its accessibility, there is an increasing body of research devoted to the use of historical AIS data to identify patterns of navigation. Much of this research is focused on the detection of anomalous patterns of movement, which is of interest to the U.S. Department of Defense and other organizations with an interest in maritime security.

The purpose of our thesis is to identify patterns of movement, and factors that affect movement, for vessels in the Gulf of Mexico near Port Fourchon, Louisiana. Located approximately 100 miles south of New Orleans, Port Fourchon has a high volume of maritime traffic that is represented in archived AIS data. Although Port Fourchon is strongly associated with the offshore oil and gas industry, it also sees activity from fishing vessels and pleasure craft. During the month of April 2014, the time period of our study, vessels in or near Port Fourchon transmitted about 200,000 AIS records per day.

An analysis of vessel movements over an extended period of time using AIS data requires substantial preparatory work. Measurements of longitude, latitude, speed, course, and time are subject to errors that often are identified only in the context of movement of a specific vessel. We develop an outlier-detection algorithm to identify and remove gross errors (“outliers”) from our analysis. Having an automated approach to outlier detection is important given the high volume of data and the larger objective to have algorithms that can operate with minimal human intervention.

We refer to a collection of time-ordered positional measurements from a single vessel as a *track*. Over the period of a month a vessel may leave and return to Port Fourchon several times. We segment each track into a series of *subtracks* consisting of a trip from Port Fourchon to a *stop point* or vice-versa, and classify a subtrack as either *outgoing* or *incoming*, respectively. Stop points are identified as locations at which the movement of a vessel is below a threshold for an extended period of time. For vessels that either depart from or return to Port Fourchon, many of the associated stop points are identified as offshore oil or gas platforms, which number in the hundreds in the Gulf of Mexico near Port Fourchon. We consider subtracks for which Port Fourchon is either a start point or an end point and the other stop point is at least 20,000 meters from Port Fourchon.

Our first objective is to examine the effectiveness of statistical clustering techniques for finding patterns of movement in the 2,712 outgoing subtracks for the month of April 2014. In much of the research literature on prediction of vessel movements, clustering plays an important role in reducing subtracks to nearly homogeneous groups based on similar motion profiles. Port Fourchon, however, does not have well-defined shipping lanes. In order to apply clustering, a matrix of distances or dissimilarities between subtracks is required, which poses several challenges. Subtracks are not synchronized; vessels maintain different velocities; and, subtracks have unequal numbers of AIS measurements. Initially, we consider only the positional attributes of the subtracks, which we convert to vectors of equal length through the use of interpolation. For interpolation we choose a set of odometer distances (5,000 meters to 30,000 meters in 5,000-meter increments) from Port Fourchon, giving us six different longitude-latitude pairs. We then calculate the distance between two tracks as the averaged Haversine distances at the interpolation points. Doing this for each pair of outgoing subtracks results in a distance matrix that has 2,712 rows and 2,712 columns, which we use for cluster analysis.

We use Partitioning Around Medoids (PAM) as a clustering method, and specify a number of clusters ranging from 2 to 10 to find the best solution. Several variations are considered, including the use of

- Positional data only;
- Positional data with weighted averaging;
- Positional data for slower vessels only;
- Positional data for faster vessels only;
- Positional data with ship type.

The combination of non-quantitative information such as ship type together with positional data requires the use of a dissimilarity measure that is appropriate for mixed-type data. We find that using positional data only with four clusters gives the best clustering solution according to a metric that is commonly used with the PAM technique. The quality of the solution is considered “reasonable” but not “strong” according to guidelines that are recommended in literature for interpretation of the metric. Graphically representing the subtracks shows that the clusters are not well separated as one would anticipate if shipping lanes were present. This finding suggests that clustering is not a reliable technique for stratifying vessel movements in and out of Port Fourchon.

Our second objective is to examine the effects of meteorological and oceanographic data on vessel movements using regression analysis. To support this analysis we combine two other data sets with the AIS data. The first is a dataset from the Bureau of Ocean Energy Management (BOEM) of the U.S. Department of the Interior which provides latitude-longitude locations of all oil and gas platforms in the Gulf of Mexico. Use of the BOEM dataset allows us to correlate stop points in the Port Fourchon area with these platforms, and to identify those that are frequently visited by single vessels. We identify fourteen vessel-platform pairs that together comprise 517 subtracks, with each pair having at least 20 subtracks. For each of the fourteen routes we derive the great-circle route, which has the shortest distance for travel between Port Fourchon and an offshore platform. For a given subtrack, its deviation from the shortest route is taken to be the average distance of its AIS positions to the great-circle route, a variable to which we refer as DISTANCE. Our aim is to assess the effects of weather and sea-state on DISTANCE.

Hourly data on weather and sea-state data is obtained from a buoy located in the Gulf of Mexico approximately 39,000 meters south of Port Fourchon, which is archived by the National Data Buoy Center of the U.S. Department of Commerce. The buoy data measurements include wind direction, wind speed, and wave height, which we use as explanatory variables with the logarithm of DISTANCE as the outcome variable. Using wind direction and the course of a vessel, we resolve wind speed into downwind (aligned with the course) and crosswind components. We also use a categorical variable that identifies the fourteen vessels as a predictor.

Linear regression of DISTANCE on the predictor variables reveals that the vessel and the downwind component are significant predictors. There is substantial variability among vessels in how closely they adhere to a great-circle route. Of the fourteen vessels considered in our analysis, their average distance from a great-circle route ranged from a few meters to nearly 3800 meters. The effect of downwind is that it is negatively associated with DISTANCE. Stated another way, the stronger the headwind (which is minus the downwind), the greater the distance. For every increase of 1 meter per second (approximately 2.24 miles per hour) of headwind, DISTANCE increases on average by approximately one percent. When strong headwind is present this effect can be substantial. This finding underscores the importance of including weather and sea-state in algorithms that are designed to predict the motion of vessels.

ACKNOWLEDGMENTS

Words cannot describe how incredibly grateful I am to my advisor, Dr. Robert A. Koyak. Professor Koyak, you truly are an amazing advisor and I cannot thank you enough for all your mentorship and guidance. I would have never made it through this process without you. Special thank you to my second reader, Dr. Jonathan K. Alt, for your help and guidance.

I would like to thank Andres Otero for being my accountability partner throughout the workings of this thesis. Your discipline and influence to stay focused and on track was unbelievably helpful and necessary to make it to the finish line.

I would also like to thank my friends and family for continued support during my time as a graduate student at NPS.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

The U.S. Navy maritime mission is to increase national security through maintaining a presence at sea and deterring dangerous opponents and adversaries. In 2007, the U.S. Navy developed the "Maritime Domain Awareness" (Department of the Navy [DON], 2007) directive charging the Surface Navy to develop better situational awareness of the global surface picture and to identify anything that could threaten the safety of the U.S. homeland (DON, 2007). One of the challenges is that the "Navy is increasingly faced with irregular opponents who employ asymmetric methods and capabilities against U.S. interests" (DON, 2007). In the maritime world this threat is heightened as the number of sea-going vessels is increasing rapidly which makes surveillance of vessels in an area of interest increasingly difficult.

A widely-used source of information on the global movement of sea-going vessels is the Automatic Identification System (AIS). AIS was launched in the year 2000 and required for use by the International Maritime Organization (IMO) for certain classes of vessels in 2002 (IMO, 2016). U.S. federal law mandates the use of AIS for vessels above a specified size and for foreign vessels operating in U.S. national waters (AIS, 2017). U.S. Department of Homeland Security requires that AIS be used for the following types of vessels (AIS, 2017):

- Self-propelled vessel 65 feet or more in length engaging in commercial services;
- Towing vessel of 26 feet or more in length with more than 600 horsepower engaged in commercial services;
- Self-propelled vessel certified to carry more than 150 passengers;
- Self-propelled vessel engaged in dredging operations that restrict the passage of other vessel traffic;
- Self-propelled vessel carrying dangerous cargo;
- Self-propelled vessel carrying flammable liquid in bulk;
- Fishing industry vessels.

AIS provides a wealth of information on nearby vessels that can aid a pilot in navigation. It includes ship name, speed, course, destination, and latitude and longitude as well as weather reports. The AIS systems used by many vessels are able to display this information graphically, an example of which is shown in Figure 1.

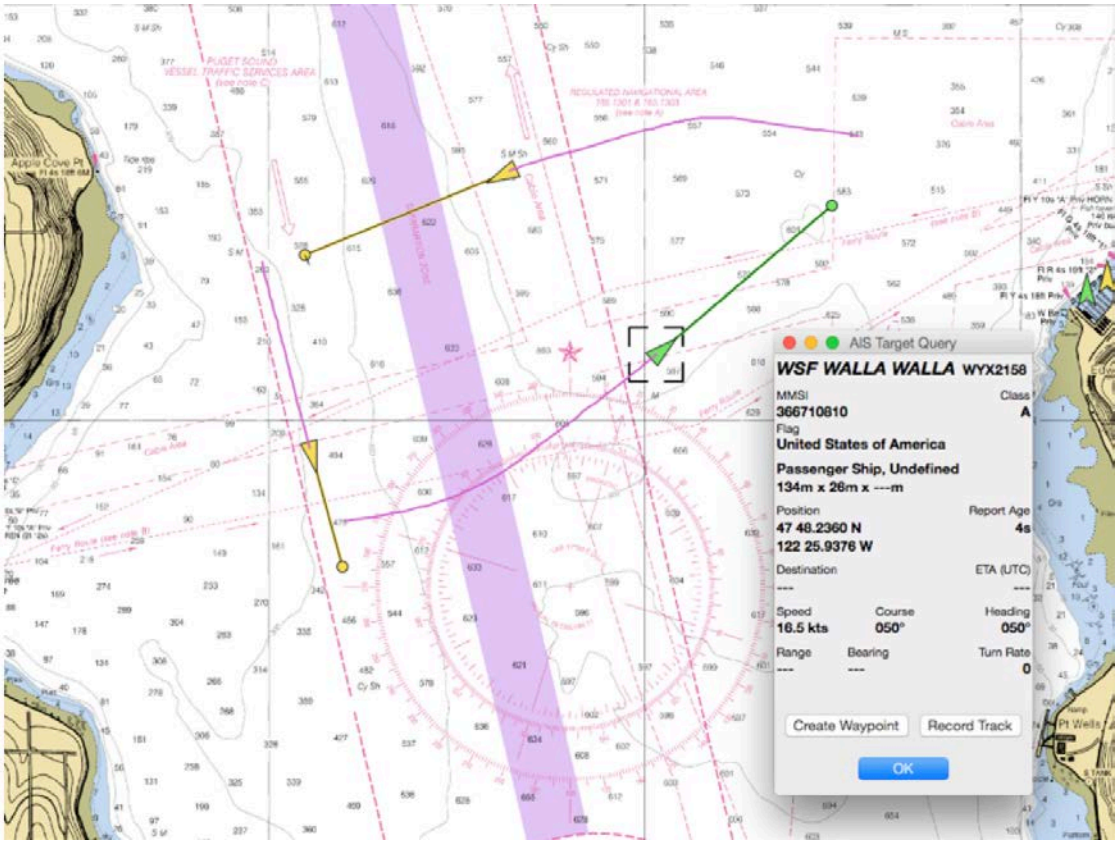
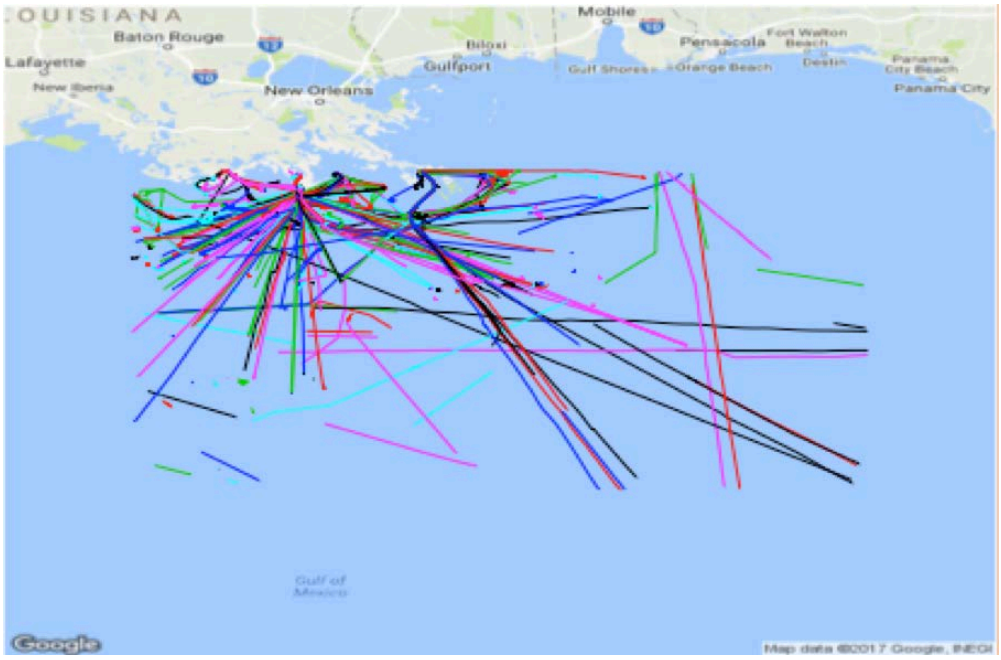


Figure 1. Graphical Output of AIS Device. Source: Burch (2016).

This display has the same layout as a paper chart on which nearby vessels are positioned. Land areas are shown in brown and coastal waters are shown in blue. The dark purple stripe is a shipping lane. Water depths in fathoms are shown as numbers. One fathom is equal to six feet. The vessels being tracked on the electronic display are the triangles in yellow and green. Dark purple lines connected to the triangles indicate the trajectories of vessels, and yellow and green lines on a triangle indicate the direction to which the vessel is traveling. A useful feature of some AIS displays is that the user may select a contact and instantly gain knowledge of the vessel. For example, in Figure 1, the

pilot selected the green vessel marked with a green triangle and the display in the gray box shows information regarding that vessel. The vessel is a U.S. passenger ship traveling at 16.5 knots in the direction of 050 which is to the northeast (50 degrees clockwise from due north). AIS also allows a pilot to broadcast information about the activity of the vessel and its destination. For example, a fishing vessel can change its status to read "engaged in fishing operations," warning others of its limited ability to abide by rules of navigation due to restricted movement caused by fishing lines and nets hanging from the vessel.

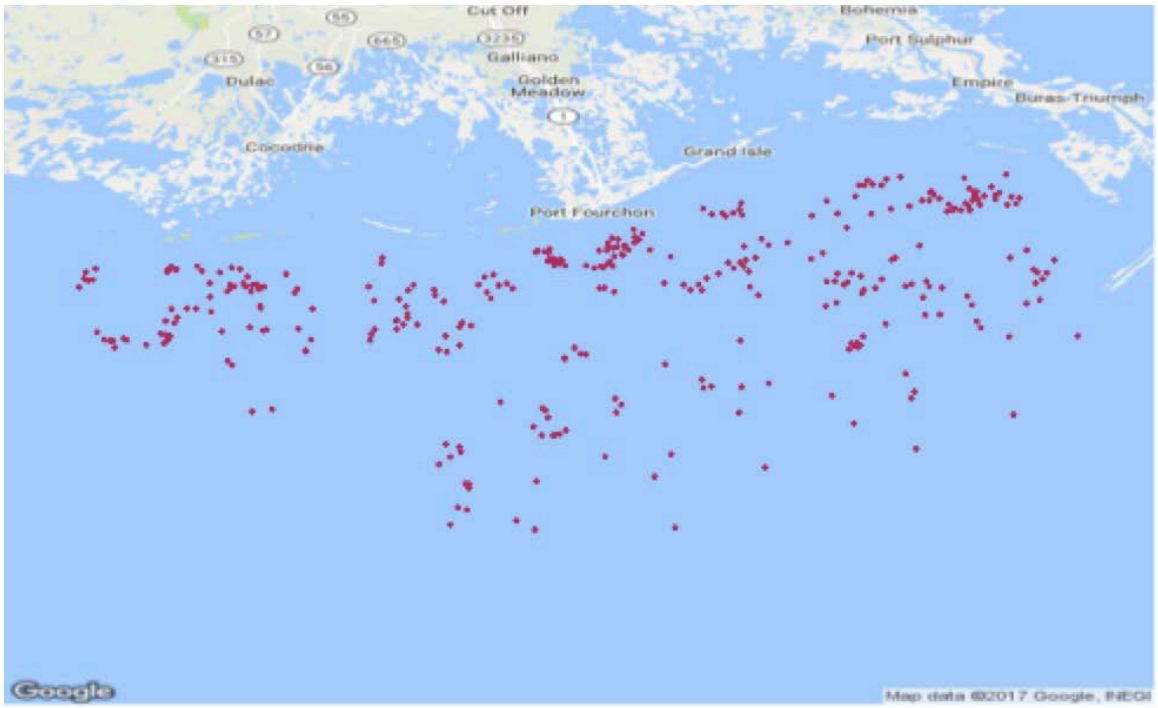
Because AIS is publicly available information, a number of parties archive its data to support analytical projects of which pattern recognition, prediction of future vessel movement, and anomaly detection are major topics. Our thesis belongs to this class of research. We examine AIS data for the month of April 2014 with the objective of describing characteristics of vessel navigation in an area of interest. Due to the world-wide use of AIS the data are voluminous. The AIS data for one day in April 2014 comprises nearly 2.5 million observations and requires approximately a half-gigabyte of storage. We limit our focus to AIS data for ships in the Gulf of Mexico to examine shipping movements. Shipping in the Gulf of Mexico produces a large amount of AIS data from cargo vessels, oil tankers, fishing and trolling vessels, and recreational vessels. In April 2014 nearly one-tenth of the AIS data for the entire world was concentrated in this region. More narrowly, we focus our research on vessel traffic in the area of Port Fourchon, Louisiana, which is situated on the Gulf of Mexico about 100 miles south of New Orleans which has about a quarter of all AIS data in the Gulf of Mexico region. Figure 2 shows a plot of 500 randomly selected tracks (a track is a series of AIS data for a specific vessel) near Port Fourchon.



Created using RgoogleMaps by RStudio.

Figure 2. Plot of 500 Tracks in the Gulf of Mexico near Port Fourchon from March 2014

Much of the maritime activity at Port Fourchon is related to servicing the hundreds of oil and gas platforms located offshore near the port. According to the Greater Lafourche Port Commission (n.d.), Port Fourchon services nearly 90 percent of domestic deepwater oil production in the United States, with nearly 600 offshore oil platforms located within 40 miles of the port. These platforms provide between 16 and 18 percent of U.S. oil production. More than 400 large supply vessels traverse Port Fourchon daily. The Louisiana Offshore Oil Platform (LOOP), located in the Gulf of Mexico about 18 miles south of Port Fourchon, is one of few deepwater ports in the Gulf of Mexico region that can accommodate Very Large Crude Carriers (VLCCs) and Ultra Large Crude Carriers (ULCCs). The LOOP supplies the United States with about 13 percent of its imported foreign oil, and is connected to about 50 percent of U.S. oil refining capacity. Figure 3 shows the locations of oil and gas platforms in the Gulf of Mexico near Port Fourchon.



Created using RgoogleMaps by RStudio.

Figure 3. Oil Platforms Located within 30 Miles of Port Fourchon

A. RESEARCH OBJECTIVES

Our research is focused on two objectives related to characterizing the movement of marine vessels in and out of Port Fourchon as measured by AIS data for the month of April 2014. The ultimate goal of analyzing data on ship movements is to develop a model or process for predicting future ship movements based on AIS information up to a given point in time about the vessel. This is a continuing research topic in the vessel tracking community, which to date has focused on areas where traffic segregates into well-defined "clusters" as one might find coming into or out of a major commercial port such as New York or Los Angeles-Long Beach. Our first objective is to determine the extent to which clustering is present in the Port Fourchon area that would allow its maritime traffic to be treated in a manner similar to that of other ports.

Our second objective is to examine the effects of weather and sea-state on the movements of a small number of vessels that make repeated trips between Port Fourchon and a common destination during the month of April 2014. We identify fourteen vessels

that meet these criteria which make a total of 517 trips to and from their destination points, and for each we calculate an average distance from the Great Circle route, which is the shortest route between two points on the surface of the earth. Treating average distance as an outcome variable, we develop a prediction model with wind speed, wind direction, and wave height obtained from hourly offshore buoy readings. Research to date has not addressed the effects of these variables on predicting ship movements.

B. ORGANIZATION OF THE THESIS

The remainder of this thesis is organized as follows. In Chapter II we review literature on research related to maritime navigation, particularly with respect to the use of AIS data. In Chapter III we explain in detail the data used in our analysis, the data processing steps used to render the data into usable form, and the analytical techniques that we use to address our study objectives. In Chapter IV we present the results of applying our approach to the Port Fourchon AIS data for the month of April 2014. We state our conclusions and identify topics for additional research in Chapter V.

II. BACKGROUND AND LITERATURE REVIEW

This chapter discusses AIS and its use in analyzing maritime navigation. This chapter also discusses other research in support of unsupervised learning of traffic patterns using historical AIS data.

A. BACKGROUND

All AIS information is unclassified and available to the public. Originally designed as a tool for collision avoidance, the uses of AIS have evolved to include fishing fleet monitoring and control, vessel traffic services, maritime security, aids to navigation, search and rescue, accident investigation, ocean current estimation, infrastructure protection, and fleet and cargo tracking (United States Coast Guard [USCG], 2016). As of 2017 there are more than 20,000 vessels and 475 shore-based stations that use AIS (AIS, 2017). AIS is both broadcasted from a vessel and received from other vessels automatically using a VHF transceiver. The information transmitted includes ship position, speed, and navigational status (USCG, 2016). The AIS transceiver updates the information in Table 1 every 2 to 10 seconds while the vessel is underway and every 3 minutes while the vessel is at anchor (USCG, 2016). Because the AIS transmittal is automated this information cannot be changed by the user.

Table 1. AIS Data Transmitted Every 2 to 10 Seconds or Every 3 Minutes. Adapted from Raymond (2016).

MMSI	**Maritime Mobile Service Identity**
Navigation status	At anchor, not under command, or underway using engines
Rate of turn	• 0 = not turning • 1…126/1…-126 = turning right/left at up to 708 degrees per minute or higher, respectively • 128 = no turn information available (default)
Speed over ground	0.1-knot resolution from 0 to 102 knots
Course over ground	Relative to true north
Latitude	-90° to 90°
Longitude	-180° to 180°
True heading	0 to 359°
Time stamp	Hour:Min:Sec in UTC format

The Maritime Mobile Service Identity (MMSI) uniquely identifies an AIS transceiver, which usually is synonymous with the vessel. Navigation status describes the current mobilization ability of a ship. The three possible values of navigational status are “at anchor,” “underway using engines,” and “not under command.” Not under command implies that the ship has run aground or has experienced some other type of casualty. Other ships in the vicinity are informed that the vessel “not under command” is unable to abide by rules of navigation.

In additional to automated reports a vessel is required to broadcast a non-automated static report every six minutes. Because these reports require user input the quality of information in them cannot be assured. Table 2 lists the data fields in the static AIS reports.

Table 2. AIS Information Transmitted Every 6 Minutes

Data Field Name	Description
IMO number	International Maritime Organization ship identification number
Radio call sign	Name given to communicate with vessel over VHF radio
Ship name	Owner-given name
Type of ship	Cargo, oiler, fishing, tug, passenger vessel, etc.
Dimensions of ship	Length from bow to stern and length from port to starboard, in meters
Location of positioning system	either aft or forward
Type of positioning system	Satellite-based or transceiver only
True heading	0-359 degrees, relative to true north.
Draught of ship	Draft below waterline, in meters
Destination	As stated by operator; not always available
ETA	Estimated time of arrival

The IMO number is a permanent vessel identification label that does not change when a ship is sold to another owner. Radio call sign is the identification used by a vessel when communicating with other vessels over a common radio frequency, and is often the same as the ship name. Using the call sign to contact a ship by VHF radio is a safe way to quickly understand the intentions of that vessel so that action may be taken such as altering course to avoid a collision. Unfortunately, this information is not always provided and can also be inaccurate. Harati-Mokhtari, Wall, Brooks, and Wang (2007) examine the accuracy of AIS static data inputs from a human factors perspective. The authors find that 6 percent of vessels reported no vessel type and 3 percent report “other” as the vessel type when one of the available choices (cargo, passenger vessel, etc.) would have been correct. In addition, 47 percent of vessels reported incorrect ship dimensions and 18 percent reported incorrect draught. Almost half of vessels (49 percent) of vessels did not report a destination.

B. LITERATURE REVIEW

There are several methods analysts have used to predict future ship movement or detect anomalies using either AIS or surveillance data from a camera. These techniques largely use different clustering methods to identify similarities in traffic patterns as a means for creating a threshold for anomaly detection.

In a Naval Postgraduate School (NPS) master's thesis, Tester (2013) explores the use of spatiotemporal clustering of vessels in a maritime domain using AIS data based on attributes such as location, speed, and time. The objective of his thesis was to determine if vessels of interest (e.g., vessels transporting illicit cargo) interacted with any other vessels over the period of observation. Tester used K-means clustering to group vessels based on proximity to one another, course, and speed. The ultimate purpose of his research was to support maritime domain awareness (MDA) efforts by developing an algorithm capable of identifying vessels exhibiting illicit behaviors, and to identify how other vessels involved in the operation interact with them.

The NPS master's thesis by McAbee (2013) investigates the use of the Hough transformation, a technique in image analysis used to identify imperfect shapes, to describe popular shipping routes in coastal waterways and the open ocean using historical AIS data. McAbee adopts a three-stage approach to this analysis in support of enhancing MDA. The first stage is to identify high density traffic areas. The second stage is to use the Hough Transformation to identify linear patterns within these high-density areas. The third stage is to define the width of the given highway. Once these highways are established, anomaly detection is performed by determining whether or not a given vessel is traveling within these highways. McAbee also explores the effects of annual seasonal patterns on maritime vessel.

Ristic, Scala, Morelande, and Gordon (2008) examine a method for anomaly detection using kernel density estimation (KDE). Anomaly detection is based on a predefined probability of false alarm, determined from historical AIS data. Using patterns developed from historical AIS data, real-time AIS data are then classified as exhibiting normal behavior or anomalous behavior based on where they were in the traffic patterns

identified. The authors then use a Gaussian sum tracking filter to aid in the prediction of future ship movement. Gaussian sum tracking uses weighted sums related to the historical movement of a vessel as a means of predicting future movement.

Morris and Trivedi (2008) investigate the detection of in-motion trajectories, not limited to the maritime domain but applicable to it, by studying motion patterns from video surveillance. They relate the problem to video surveillance in a parking garage and demonstrate how a constant collection of tracks can be used to predict future movement. The first stage of their approach consists of defining points of interest where interesting events occur and the second stage is to define activity paths that characterize how objects move between points of interest. The authors conclude that a vocabulary for analyzing a scene can be developed in an unsupervised fashion using historical data on the movements of objects. This vocabulary would allow for classification of past and current activity, detection of abnormal activities, prediction of future activities, and characterization of interactions between objects.

In a subsequent paper, Morris and Trivedi (2011) continue discussion on the use of video surveillance to develop a data-based method for predicting future movements and detecting anomalies. Their second paper addresses the use of Gaussian mixture modeling to connect routes through trajectory clustering and spatio-temporal dynamics of activities encoded using hidden Markov models. Similar to their earlier work, the authors build a vocabulary by identifying recurrent patterns in data. They adopt a three-stage hierarchical learning process for creating the vocabulary and predicting behavior. The first stage uses Gaussian mixture modeling to discover nodes and points of interest. The second stage uses trajectory clustering and spatio-temporal dynamics to determine number of activities in a scene and learn similarities in routes. The third stage uses a Hidden Markov Model to make future route predictions and to estimate probabilities of anomalies.

In his doctoral dissertation, Laxhammar (2014) discusses several approaches to anomaly detection in the maritime domain using historical AIS data. The output of these approaches all use a predetermined threshold for identifying anomalies, where if the value of the given AIS contact is above the threshold then the contact is labeled an

anomaly. Laxhammar provides the reader with the advantages and disadvantage of each approach by discussing computational efficiency of each algorithm and which algorithms work better with small versus large datasets.

III. METHODOLOGY

In this chapter, we explain the process that we use to prepare the AIS data for analysis, and the techniques that we use to address the study questions posed in Chapter I.

A. DATA INGESTION AND INITIAL PROCESSING

The AIS data is received in AIVDM/AIVDO format, which is a method of collecting and integrating information from publicly available sources (Raymond, 2016). Initially, the AIS transmitters broadcast their positions from vessels, navigation markers, and shore positions. When the data is received, it is in the form of a text packet composed of bit character strings that require conversion to a useable format. Figure 4 is an example of what an AIVDM/AIVDO data packet looks like.

```
!AIVDM,1,1,,B,177KQJ5000G?tO`K>RA1wUbN0TKH,0*5C
```

Figure 4. Visual of an AIVDM/AIVDO Data Packet. Source: Raymond (2016).

In this example, there are a total of 7 fields, each separated by commas, that provide information to the interpreter on the contents of the packet payload (Raymond, 2016). We will focus on field 6 (“177KQJ5000G?tO`K>RA1wUbN0TKH”), which is the actual AIS information we want. All characters are converted to a bit format and concatenated, forming the binary payload of the sentence. The bit strings are segmented and then converted back to alphanumeric data.

The first section of interpreting the message payload is the message type, which is the first 6 bits in the character string. Table 3 shows the most popular message types.

Table 3. Popular Message Types. Adapted from Raymond (2016).

01	Position Report Class A
02	Position Report Class A (assigned schedule)
03	Position Report Class A (response to interrogation)
05	Static and Voyage Related Data

In accordance with IMO standards, message types 1, 2, or 3 are 168 bits long and are updated every 2 to 10 seconds while a vessel is underway and every 3 minutes when a vessel is at anchor (IMO, 2016). This is the automated information that is transmitted without intervention by the vessel crew or other persons. Figure 5 shows how each six-bit segment is converted for message types 1, 2, and 3.

Field	Len	Description	Member
0-5	6	Message Type	type
6-7	2	Repeat Indicator	repeat
8-37	30	MMSI	mmsi
38-51	14	Year (UTC)	year
52-55	4	Month (UTC)	month
56-60	5	Day (UTC)	day
61-65	5	Hour (UTC)	hour
66-71	6	Minute (UTC)	minute
72-77	6	Second (UTC)	second
78-78	1	Fix quality	accuracy
79-106	28	Longitude	lon
107-133	27	Latitude	lat

Figure 5. Example of Information in Message Types 1, 2, or 3. Adapted from Raymond (2016).

Message type 5 is the static information transmitted by a ship and is done so every 6 to 10 minutes. This report is 424 bits and an example of this report is in Figure 6.

Field	Len	Description	Member/Type
0-5	6	Message Type	type
6-7	2	Repeat Indicator	repeat
8-37	30	MMSI	mmsi
38-39	2	AIS Version	ais_version
40-69	30	IMO Number	imo
70-111	42	Call Sign	callsign
112-231	120	Vessel Name	shipname
232-239	8	Ship Type	shiptype
240-248	9	Dimension to Bow	to_bow
249-257	9	Dimension to Stern	to_stern
258-263	6	Dimension to Port	to_port
264-269	6	Dimension to Starboard	to_starboard
270-273	4	Position Fix Type	epfd
274-277	4	ETA month (UTC)	month
278-282	5	ETA day (UTC)	day
283-287	5	ETA hour (UTC)	hour
288-293	6	ETA minute (UTC)	minute

Figure 6. Example of Message Type 5. Adapted from Raymond (2016).

The parsed AIS data are separated into positional and static reports and stored in a convenient format for retrieval. To support our research, the Center for Multi-Int Studies (CMIS) at NPS provided us all the AIS reports world-wide for the calendar year 2014. Additionally, data for January through April 2014 were parsed into comma-separated value (CSV) files by SPAWAR (U.S. Navy) and provided to us by CMIS. We focus our effort on data from the parsed files, comprising the first four calendar months of 2014. Daily positional report files are converted to SpatialPointsDataFrame format using R package sp (Bivand, Pebesma, & Gomez-Rubio, 2005). This format is particularly

useful for the analysis and graphical display of geographical data. Daily static reports are converted to R data frame format. Both positional and static data are sorted in chronological order and duplicate records are removed. For the thirty-day period comprising the month of April 2014, the AIS positional data set contains 7,414,423 records limited to the range (25.0, 31.0) latitude and (-93.0, -87.0) longitude. This rectangular region is in the U.S. coastal region of the Gulf of Mexico including the entire Louisiana coastline. The same region and time period has 88,474 static AIS records.

B. CONVERTING DATA TO USEABLE FORM

An individual AIS positional report provides information about the location and movement of a vessel, identified by its transceiver (MMSI) handle, at a particular point in time. When these reports are linked together, they provide a historical record of where the vessel has been, the routes that it followed to its various ports of visit, and its motion characteristics (e.g., velocity). Thus, over a given period of time, the AIS data are properly considered to be a collection of tens of thousands of vector-valued time series that are observed asynchronously. A typical time series, that we call a *track*, is represented as a collection of measurements $(x_i, y_i, \mathbf{z}_i, t_i), i = 1, \ldots, n$ where t_i is the time stamp of the i^{th} observation; x_i and y_i are the longitude and latitude reported at time t_i; and $\mathbf{z}_i$ is a vector of other variables observed at time t_i

Variables in $\mathbf{z}_i$ include not only time-stamped attributes such as speed and course that are obtained from the positional reports, but also items such as call sign, ship name, ship type (e.g., cargo ship), ship dimensions, and the destination, all of which are information from the static AIS report. Unlike the positional reports, AIS static reports are manually reported, and as such their accuracy cannot be assured. For simplicity, we drop reference to $\mathbf{z}_i$ and let the position and time measurements $(x_i, y_i, t_i), i = 1, \ldots, n$ denote a track.

C. OUTLIER DETECTION

The AIS data are subject to measurement errors in positions, motion characteristics, and time stamps. Of concern are gross errors of a magnitude that can distort a statistical analysis. Some of these errors are obvious, such as a longitude or latitude that falls outside of an allowable range (–180 to 180 degrees of longitude and –90 to 90 degrees for latitude); but more typically, gross errors are discovered in the context of a track. A position-time measurement (x_i, y_i, t_i) is unusual relative to the preceding track measurement $(x_{i-1}, y_{i-1}, t_{i-1})$ if the distance traveled over the time interval $[t_{i-1}, t_i]$ is inconceivable for the vessel given its speed limitations. A displacement that would require an average speed of 80 knots sustained for a ten-minute period, for example, would not be possible for an oil tanker. The question is which of the two measurements is the potential outlier? We answer this question by calculating the average speed of the vessel between two measured locations by dividing the distance by the time increment. An outlier typically is unusual relative to all or nearly all other measurements using the average velocity criterion with an appropriate threshold.

We mention that distance between two positions in longitude-latitude coordinates is calculated using Haversine distance which is the great circle arc length based on a spherical Earth model, as implemented in the function distHaversine provided in the R package geosphere (Hijmans, 2015). The formula for the Haversine distance is given by the following formula:

$$d((x_1, y_1),(x_2, y_2)) = 2r_0\sqrt{\sin^2\left(\frac{y_2 - y_1}{2}\right) + \cos(y_1)\cos(y_2)\sin^2\left(\frac{x_2 - x_1}{2}\right)},$$

where $r_0 = 6{,}378{,}137$ is the approximate radius of the Earth, in meters. A more accurate calculation may be obtained using an ellipsoidal Earth model but the computational time is substantially increased while the improvement in accuracy is small for the geographical range that we consider in this thesis.

To detect potential outliers in a track we calculate the speeds (distances divided by the absolute values of time differences) for all pairs of observations for which the

absolute time difference exceeds a threshold (e.g., one minute). For each observation, the number of exceedances of a speed threshold (e.g., 2,000 meters per minute, which is equivalent to about 75 miles per hour) is calculated. The observation with the largest nonzero count is flagged as a potential outlier and set aside. The total number of exceedances is then recalculated and the process is repeated until none of the remaining observations have any exceedances. Setting a minimum time threshold is necessary to rule out designating outliers in observations with extremely large speed ratios which arise due to small time differences combined with measurement error in both AIS position and time reports.

D. TRACK SEGMENTATION

Over the course of time a vessel makes multiple trips to and from ports of call or other locations that we define as *stop points* where the measured speed of the vessel remains small (essentially zero) for a period of time. Learning the stop points of a vessel is essential to understanding its movement patterns. For example, we can learn from a cargo vessel that makes frequent stops to only two different Shell oil platforms? Also, it is a first step in a process of segmentation of tracks into *subtracks*, which are movements between pairs of stop points. In this thesis, we focus on subtracks for which Port Fourchon is either a start point or an end point. We first identify tracks that have AIS observations that are within 5000 meters of Haversine distances of Port Fourchon (longitude = –90.19444, latitude = 29.10556) and use this criterion to imply being at that location. For these tracks we then locate periods of travel between periods of being stopped at Port Fourchon. The first stop point either coming into or leaving the port is found and the subtrack is ended at that point. For stopping, we calculate a smoothed vessel speed by measuring total distance traveled in a time window around a given AIS-observed time, and divide by the time difference. Although AIS reports include vessel speed we have found it less reliable to use these measurements than to use smoothing. Our criterion for stopping is that a vessel moved less than 500 m over a period of at least 20 minutes. There are several possibilities for identifying the endpoints of a subtrack:

1. The origination point of the incoming subtrack is unknown since it truncated from the left side of the end point.

2. On the contrary, the destination of the outgoing track is unknown since it is truncated from the right side of the end point.

3. Outgoing subtracks reach a point where the speed of the vessel becomes very slow (e.g., stopped). This point is the destination.

4. Working backwards from Port Fourchon, an incoming subtack reaches a point of low speed which is considered the origination point.

5. Between two consecutive time periods in which a vessel is at Port Fourchon, the vessel goes out and then returns, but there is no detectable low-speed point. In this case, two subtracks are defined. The point of maximum distance from Port Fourchon is taken to be the destination for an outgoing subtrack, and also the point of origin for an incoming subtack.

For our analysis we consider subtracks that have Port Fourchon either as an origination or destination point, and for which the maximum distance from Port Fourchon is at least 20,000 meters. This reduces the positional AIS data to 1,775,071 records and the static report file to 23,177 records, comprising 730 tracks (distinct MMSI values) that have some association with either going to or from Port Fourchon. These tracks are compressed into 8,906 subtracks of which 4,469 are outgoing and 4,439 are incoming. Associated with these subtracks are 8,765 stop points, many of which are redundant.

Figure 7 depicts a randomly selected track corresponding to one of the 730 vessels in the April 2014 data. The MMSI of the vessel is shown as the title. The vertical axis is distance from Port Fourchon in kilometers. There are five subtracks in this plot. The blue subtracks represent the vessel moving towards Port Fourchon and the pink subtracks represent the vessel moving away from Port Fourchon. The green horizontal line is the 5,000-meter threshold below which a vessel is regarded as being at Port Fourchon. Of the five stop points, four are within 650 m of each other, and the other is about 2,200 m away from these four. It is likely that the four are the same stop point, and it is possible that the other is as well.

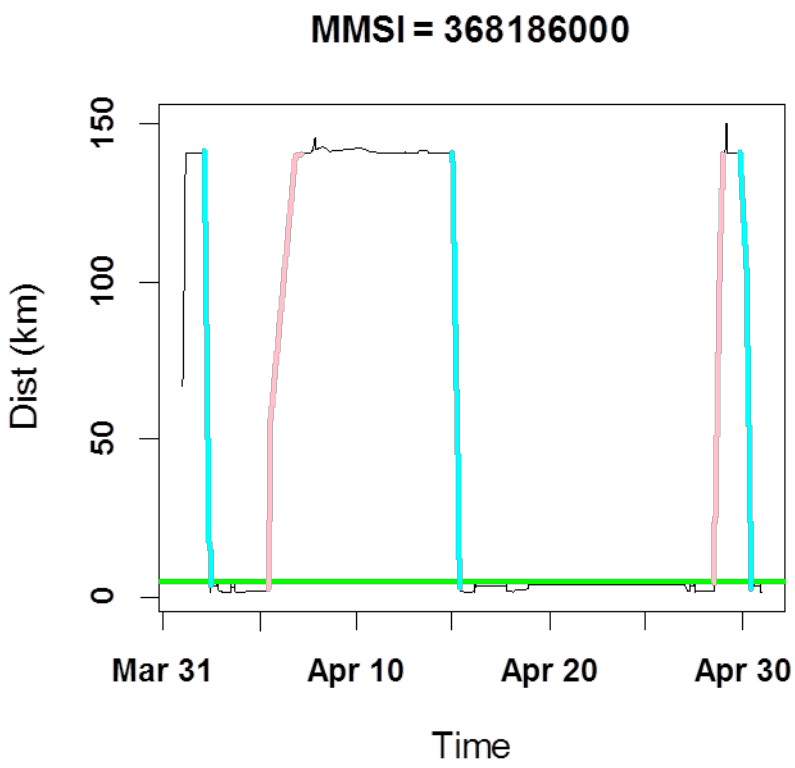


Figure 7. Sample Monthly Voyage from a Randomly Selected MMSI

Applying the outlier detection methodology outlined above to the Port Fourchon subtracks, we identify 1,008 observations as potential outliers out of a total of 178,546 records, which is less than 0.6 percent of the total. These cases arise due to errors in positional measurements or in time stamps. We examine these observations by considering their maximum distance from other observations in the same track that are time-stamped to within an hour of the potential outlier in question. Of the 1,008 potential outliers, over half (564) have maximum distances less than 46,300 m, which correspond to speeds less than 25 knots. Because a speed of 25 knots is not unusual for sea-going vessels, it implies that the problem in these cases may be with the time stamps. Only eleven potential outliers have maximum distances that exceed 92,600 m (50 knots) and only one case exceeds 185,200 m (100 knots), which plausibly arise from positional errors.

E. SUBTRACK ALIGNMENT

A primary objective of an analysis of AIS data is to identify subtracks that are similar to each other. Groups of subtracks that have similar characteristics are used to develop predictions of movement that apply to those groups separately. Subtracks, however, are offset from each other with respect to time and do not contain the same number of measurements. Familiar vector-based metrics (e.g., Euclidean) therefore cannot be used to measure the distance of one subtrack from another.

To calculate positional distance between two subtracks with possibly different speeds we first convert the subtracks to sequences of equal length using interpolation. For outgoing subtracks we approximate the positions of the subtracks when they have logged a set of prescribed distances after having left Port Fourchon. This also removes the time element from consideration. For example, a set of distances starting at 5,000 m and ending at 30,000 m with increments of 5,000 m has six interpolation points. We find bracketing AIS observations and use simple linear interpolation to approximate the positions of a vessel at these distances. If bracketing AIS observations cannot be found at a particular distance, missing values are recorded for the interpolated position. For incoming subtracks the procedure is used in a similar manner. The distance between two subtracks is taken to be the averaged Haversine distances at the interpolation points. Although we use non-weighted averaging in our analysis, a weighted average may be used if desired; i.e., to give more influence to subtrack differences when the vessels are close to Port Fourchon.

F. STOP POINT SMOOTHING

The same stop point visited by two different subtracks will not have identical positions due to measurement error and inaccurate stop point estimation using the segmentation algorithm. To reduce redundancy we process the stop points so that those which are close to each other have their coordinates replaced by averaged values. Using a distance of 1,000 meters as a threshold for closeness, smoothing in the described manner reduces the number of stop points from 8,765 to 1,459. Increasing the distance threshold to 2,000 m reduces the number of stop points further to 1,051. This is not surprising,

considering that many of these stop points are oil and gas rigs, hundreds of which are densely situated offshore in close proximity to Port Fourchon. Other stop points may consist of fishing grounds or other service locations (e.g., pipelines) associated with the oil and gas industry.

G. OTHER DATA SETS USED

We use two other data sets to aid in our analysis of the AIS data from the Port Fourchon area.

1. Oil and Gas Platforms in the Gulf of Mexico

The Bureau of Ocean Energy Management (BOEM) of the U.S. Department of the Interior makes publicly available extensive data on 6,364 oil and gas platforms in the Gulf of Mexico through its website at https://www.data.boem.gov/homepg/data_center/platform/platform.asp. This website provides the coordinates of the platforms and detailed information about ownership, size of operation, and other platform attributes. We use the BOEM data to identify subtrack stop points that are associated with oil and gas platforms.

2. Meteorology and Oceanography Data

The National Buoy Data Center (NBDC) of the U.S. Department of Commerce maintains extensive archives of weather and sea-state data collected at buoys in U.S. coastal waters. We use data from buoy SPLL1 at South Timbalier Block 52, which is owned and maintained by Louisiana State University. SPLL1 is located at –90.483 degrees longitude and 28.867 degrees latitude, approximately 39 km south of Port Fourchon. Hourly data from SPLL1 are available for all of the year 2014, which include wind direction, wind speed, gust speed, and wave height. By associating the time stamp from the SPLL1 data with the AIS data, we can gauge what the weather and sea-state conditions were to support our regression modeling effort that we describe as follows.

H. CLUSTER ANALYSIS

The analysis in this thesis largely focuses on the Partitioning Around Medoids (PAM) clustering technique, for which a set of multivariate objects are partitioned into *k* clusters centered on *medoids* which have properties similar to medians (Kaufman and Rousseeuw, 1990). The goal of PAM is to define the medoids to be distinct observations with subscripts $m_1, m_2, \ldots, m_k$ that minimize the following objective function:

$$g(m_1, m_2, \ldots, m_k) = \sum_{i=1}^{n} \min_{t=1,\ldots,k} d(i, m_t)$$

where $d(i, j)$ is a distance or dissimilarity measure for comparing observations i and j across attributes (variables) that may be quantitative, non-quantitative, or a mix of both types. The number of clusters, *k*, must be specified. Dissimilarities are non-negative numbers that are small when *i* and *j* are near each other and large when they are far apart. For our analysis, we deal with quantitative variables such as speed, course, and positional coordinates; and qualitative variables such as ship type.

The strength of association of an object with its assigned cluster is measured using the silhouette value, which is defined in Kaufman and Rousseeuw (1990). The silhouette value is a number between –1 and +1 where –1 signifies poor association and +1 signifies strong association. When clusters are perfectly separated by linear surfaces the silhouette values tend to the upper end of the scale. Likewise, when clusters are not present the silhouette values tend to average near zero, with both positive and negative values. Silhouette values averaged by cluster give a measure of separation of the individual clusters. The silhouette coefficient, which is the average of silhouette values over the entire set of observations, measures the overall quality of the cluster solution. Values of the silhouette coefficient that are greater than 0.5 indicate a "reasonable structure" in the taxonomy of Struyf, Hubert, and Rousseeuw (1997), shown in Table 4.

Table 4. Silhouette Coefficient Indicator. Adapted from Struyf et al. (1997).

Silhouette Coefficient	**Interpretation**
0.71 – 1.00	Indicates a strong structure
0.51 – 0.70	A reasonable structure
0.26 – 0.50	Structure is weak
≤ 0.25	No substantial structure found

In R, the package cluster implements PAM clustering and provides useful graphics based on the silhouette values (Maechler, Rousseeuw, Struyf, Hubert, & Hornik, 2016). A silhouette plot shows the observations, ordered by their assigned clusters on the vertical axis, with silhouette values on the horizontal axis. When strong clustering is present the silhouette plot for each cluster exhibits a rectangular shape, with large gaps between clusters, and few or no negative values. Figure 8 shows silhouette plots for subtracks going out of Port Fourchon during a one-week period in 2014. The plot on the left uses $k = 4$ clusters for which a silhouette coefficient of 0.54 is achieved, indicating reasonable structure. The plot on the right uses $k = 8$ clusters with a silhouette coefficient of 0.43, which indicates weak structure. Clusters in the $k = 4$ solution are better separated visually than in the $k = 8$ solution.

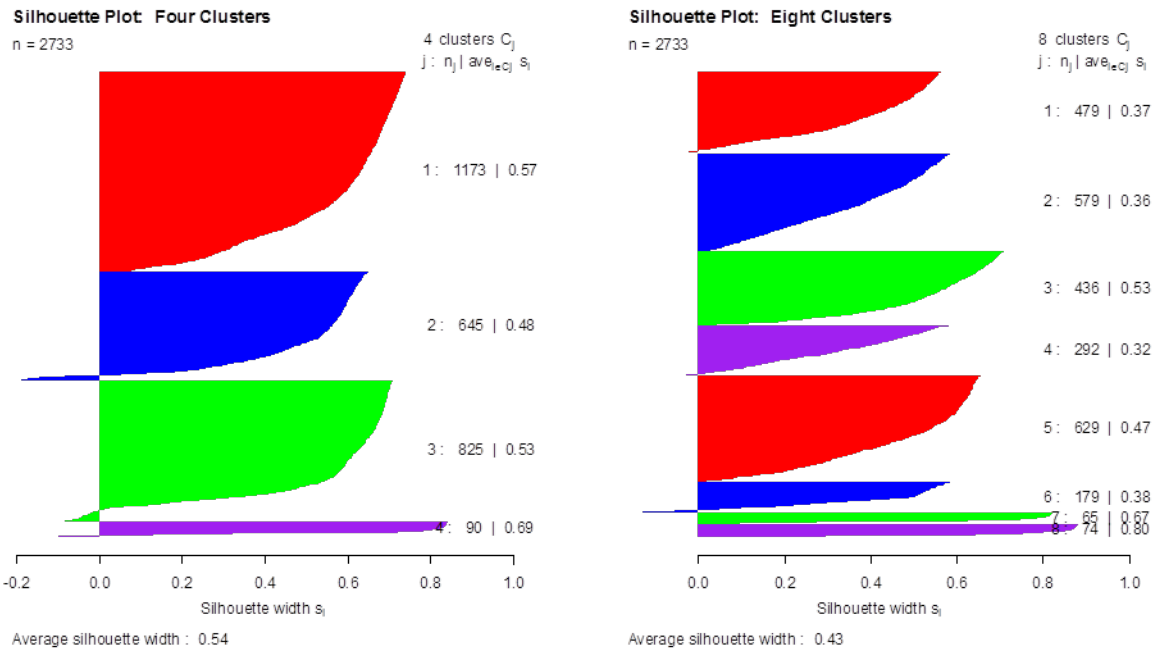


Figure 8. Example of Silhouette Plot

A shortcoming of the silhouette coefficient is that it can mask information about the strengths of the individual clusters. In the $k = 4$ solution clusters 1 and 4 fare the best at finding the right cluster for observations assigned to them with average silhouette values of 0.57 and 0.69 respectively. Clusters 2 and 3 do not fare as well with a number of observations exhibiting negative silhouette values.

One strategy for using PAM as a clustering technique is to choose the value of k that maximizes the silhouette coefficient, which can be done by examining the silhouette coefficients obtained for k over a range of values. We apply this strategy in the present example using values of k that range from 2 to 10. The best solution obtained is with $k = 4$ clusters as shown in Figure 9. In this example PAM is applied to Port Fourchon outgoing subtracks using latitude and longitude as the only variables for clustering. Distances between subtracks are Haversine distances averaged over interpolated coordinates at distances ranging from 5,000 m to 30,000 m in increments of 5,000 m.

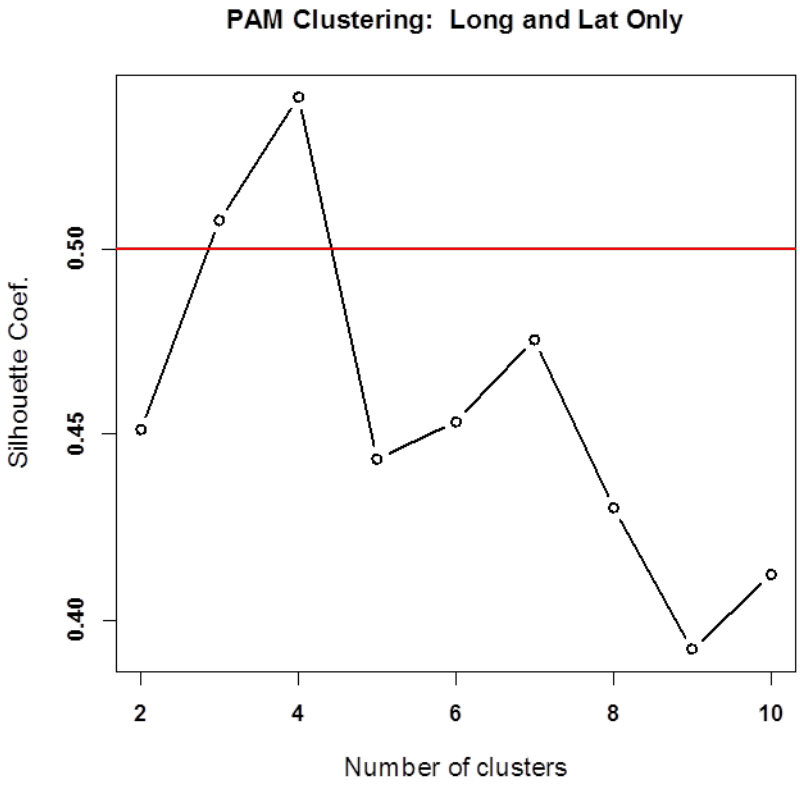


Figure 9. Example of PAM Clustering with Identified Clusters and Associated Silhouette Coefficient

I. REGRESSION ANALYSIS

A second analytical technique we use for our analysis is the use of linear regression. The purpose of regression is to explain how a response variable Y can be predicted by the values of predictor variables $X_1,\ldots,X_{p-1}$, where $p-1$ is the number of predictor variables (Faraway, 2015). There are two main objectives of regression analysis: to predict future outcomes based on given values of the predictor variables; and, to examine the interactions between the response variable and the predictor variables (Faraway, 2015). The model for predicting the value of the response variable from the predictor variables is shown below. Here, Y is the response variable, β_0 is the intercept term, $\beta_1,\ldots,\beta_{p-1}$ are slope parameters, and ε is a random error term (Faraway, 2015).

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon$$

An important aspect of regression analysis is to select a subset of the predictor variables that is useful for explaining *Y* without overfitting the noise in the model. This is often done by minimizing the sum of squared residuals or by maximizing the likelihood, together with a penalty for including an additional predictor variable. Two critical methods that we use in our analysis are the Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC), (Faraway, 2015). These two criterion based procedures are used in our analysis to indicate which predictor variables should be included in the model using stepwise-selection procedure. For both AIC and BIC, the lower AIC and BIC output number is associated with the best number of predictor variables to use for the regression model (Faraway, 2015).

It is important to conduct a diagnostic analysis of any regression model that is fit to data. The assumptions of the regression model can be violated in a number of ways such as

- Nonlinear relationships between the outcome variable and its predictors;
- Inclusion of improper predictors and exclusion of important predictors;
- Non-normality of the random error term;
- Unequal variance of the error terms.

In Chapter IV we present the results of a diagnostic analysis applied to a regression model that we develop to predict subtrack deviations from shortest paths using weather and sea-state variables as predictors.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. ANALYSIS

The analysis consists of two parts. Both parts evaluate historical AIS data for the Port Fourchon area from April 1–30, 2014. The first part of the analysis covers clustering using the PAM technique to examine how clustering works using the following factors:

- Positional data
- Positional data with weighted measures
- Positional data with slow vessels
- Positional data with fast vessels
- Positional data with ship type

We also conduct a cluster analysis of frequently encountered stop points for vessels outgoing from Port Fourchon to further explore how these subtracks may be grouped into clusters. In the second part of the analysis we conduct a regression analysis to examine how variables related to weather and sea state influence the movement of vessels that frequently travel to and from a small set of stop points that we identify as offshore oil or gas platforms.

A. PREPARATION OF DATA FOR CLUSTER ANALYSIS

As explained in Chapter III, AIS data are obtained from two sources. AIS positional data are automatically transmitted by a vessel to report its time, position, speed, heading, and other motion-related attributes. We format the April 2014 positional data as a spatial points data frame in R with 1,775,071 observations. AIS static data are manually transmitted data that report attributes of the vessel, which we format as a data frame in R with 23,177 observations. AIS positional data contain only tracks for vessels that come within 5,000 meters of Port Fourchon at some point during the month. We segment the AIS data into subtracks that are either coming into or going out of Port Fourchon. We also remove outliers that we identify as discussed in Chapter III. There are 8,906 subtracks of which 4,467 are incoming and 4,439 are outgoing.

For clustering we focus on the subset of 4,439 outgoing subtracks, which are separate time series with lengths that are unequal. We resolve the subtracks into vectors of equal length by interpolating their longitude and latitude coordinates to a common set of six odometer distances ranging from 5,000 to 30,000 meters in increments of 5,000 meters from Port Fourchon. We then match this information to the static data set to obtain the reported ship type and the ship dimensions (distance from bow to stern and from port to starboard, in meters). We take the product of the bow-to-stern and the port-to-starboard lengths as a proxy for the size of the vessel:

$$Ship\ size = (bow + stern) \times (port + starboard)$$

In the outgoing subtracks there are 617 missing values for ship size due to the required dimensions either not being reported or having an invalid entry in the static data for that vessel. We again emphasize that the AIS static reports are not quality controlled, and we have encountered instances in which the ship dimensions are clearly intended to be measured in units of feet although the instructions call for the use of meters. Our final data set consists of the following variables: MMSI, ship type, ship area, subtrack, latitude, longitude, speed, and heading interpolated at odometer distances of 5,000 to 30,000 meters in increments of 5,000 meters.

A breakdown of ship type is shown in Table 5. The most frequent category is cargo ship with 1,459 outgoing subtracks. The category "Other" captures all ship types not otherwise indicated, include those cases in which ship type is missing or invalid in the static reports.

Table 5. Ship Types for April 2014 AIS Data

Ship Type	**Number of Subtracks**
Cargo ship	1459
Other	1057
Passenger ship	780
Vessel	355
WIG	343
Tug	260
HSC	112

B. CLUSTER ANALYSIS USING POSITIONAL DATA

We initially apply cluster analysis using PAM with positional data only. When we interpolate subtrack positional data at distances from 5,000 to 30,000 meters in increments of 5,000 meters, missing values occur when a subtrack does not achieve an odometer distance of at least 30,000 m. Eliminating these cases leaves 2,712 subtracks for use in a cluster analysis. We examine how well the PAM solutions perform for the number of clusters ranging from 2 to 10. For this analysis we calculate a dissimilarity measure D between a pair of subtracks by averaging the Haversine distances of their coordinates at the interpolation points:

$$\text{Subtrack } A\text{: } \mathbf{x}_A = \left\{(x_{A,i}, y_{A,i}),\ i = 1,\ldots,r\right\}$$

$$\text{Subtrack } B\text{: } \mathbf{x}_B = \left\{(x_{B,i}, y_{B,i}),\ i = 1,\ldots,r\right\}$$

$$D(\mathbf{x}_A, \mathbf{x}_B) = \frac{1}{r}\sum_{i=1}^{r} d\left((x_{A,i}, y_{A,i}), (x_{B,i}, y_{B,i})\right)$$

Here, we use $r = 6$ and the index i indicates the interpolation point. Taken over all pairs of subtracks the resulting inter-subtrack dissimilarity matrix **D** has 2,712 rows and 2,712 columns. Figure 10 shows the results of varying the number of clusters, and indicates that the solution with $k = 4$ clusters maximizes the silhouette coefficient, which is classified as "reasonable structure" according to Struyf (1997). None of the other choices of k meet this criterion. The silhouette plot for the $k = 4$ indicates that the second cluster tends to be somewhat weaker than the others.

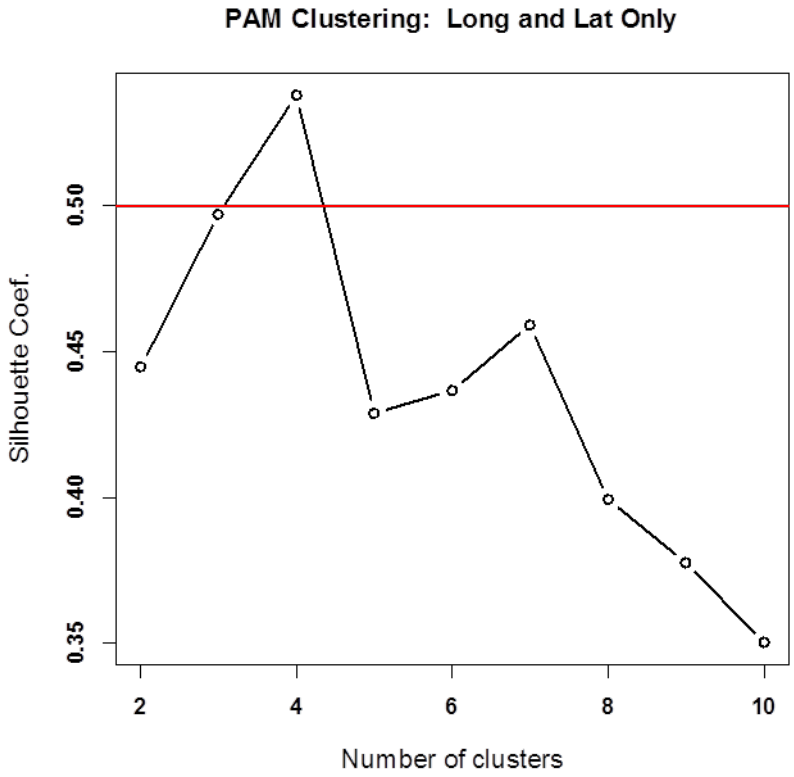


Figure 10. Clustering Results Using Positional Data Only

Figure 11 is a silhouette plot displaying how each cluster performed using four clusters as indicated in Figure 10. To the right of each cluster there are two numbers. The first is the number of subtracks in the given cluster and the second is the silhouette coefficient, explaining how well the model did at grouping subtracks into the given cluster. The silhouette coefficients in Figure 11 range from 0.48 to 0.57. The blue cluster has a silhouette coefficient of 0.48, indicating a weak structure, which is most likely due to the fact that some of this cluster produces negative results. This means that some of the clusters were improperly placed into the blue cluster. The other three clusters meet the threshold of being considered a reasonable structure.

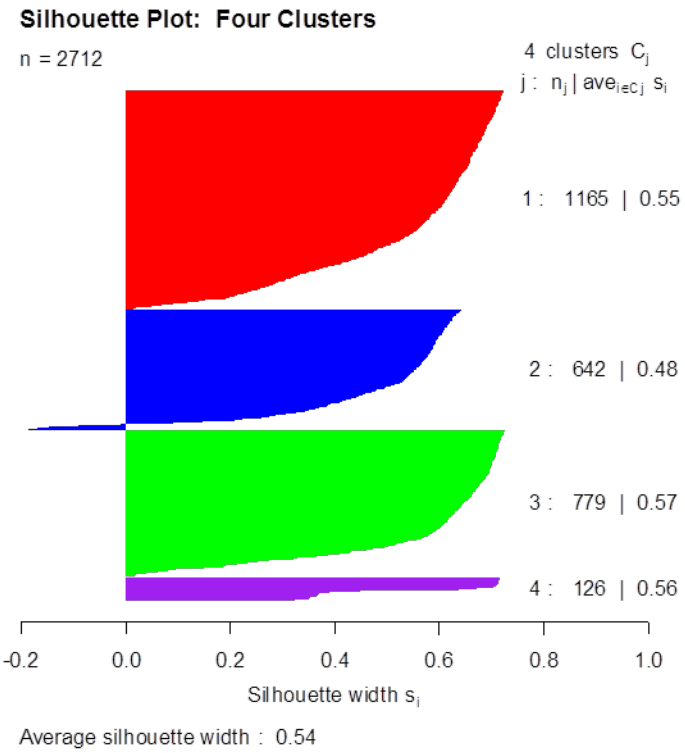
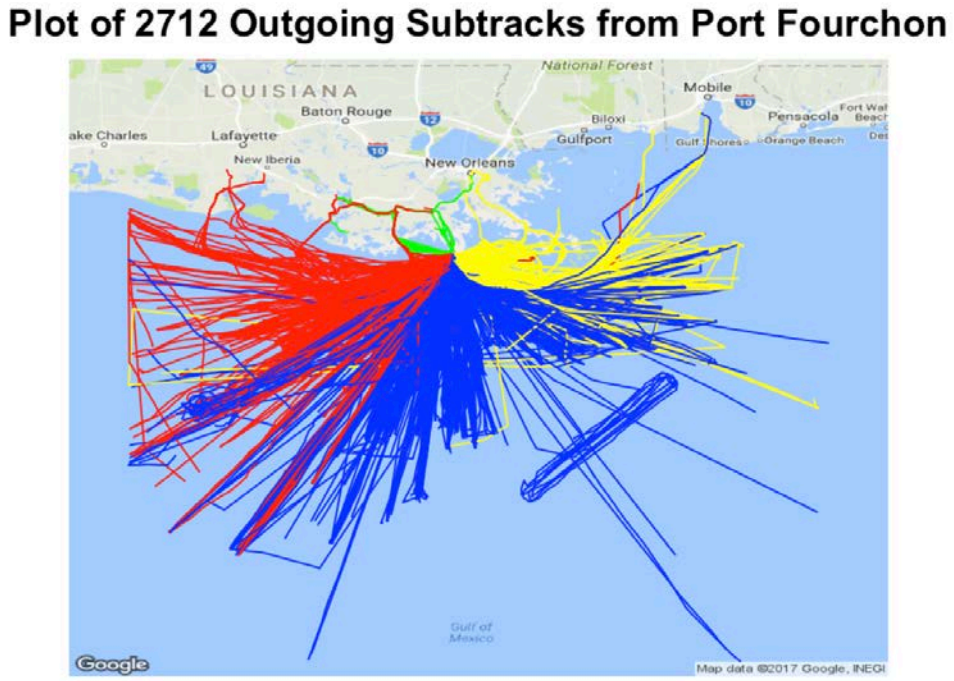


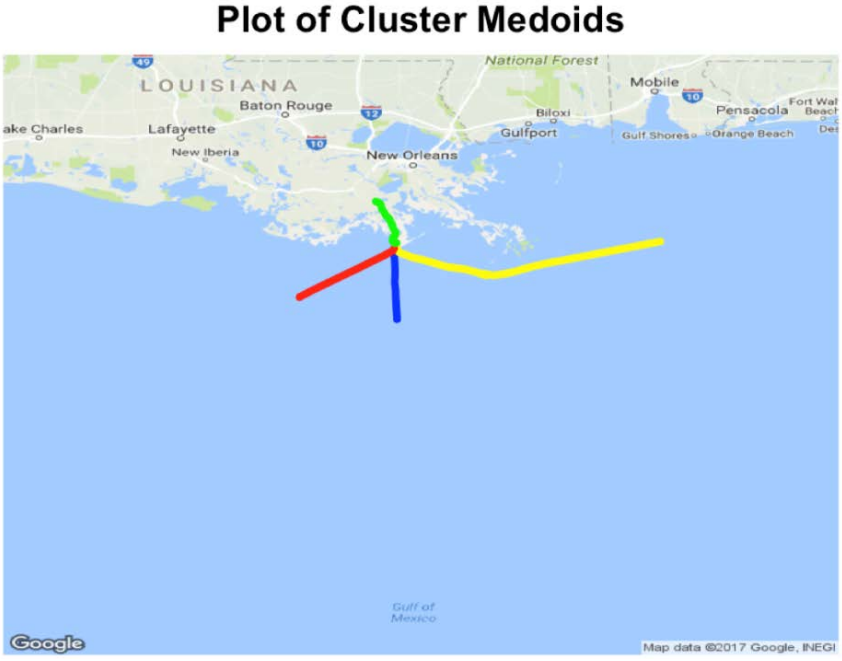
Figure 11. Clustering Results Using Positional Data Only

Figure 12 shows a plot of the outgoing subtracks with different colors indicating cluster memberships for the $k = 4$ solution. The red-coded cluster contains tracks that move to the southwest of Port Fourchon, the blue-coded cluster moves south, the small green-coded cluster moves north, and the yellow-coded cluster moves east. Figure 13 shows a plot of the cluster medoids which are "central" subtracks for each of the clusters. Although separation is present the boundaries of the clusters are not sharply defined. Because the intended use of clustering is to segregate the subtracks into a small number of relatively homogeneous navigation routes to support further analyses such as prediction and anomaly detection, it is doubtful that clustering of subtracks with Port Fourchon as a point of origination or destination will benefit for the use of this technique.



Created using RgoogleMaps by RStudio.

Figure 12. Plot of 2,712 Outgoing Subtracks



Created using RgoogleMaps by RStudio.

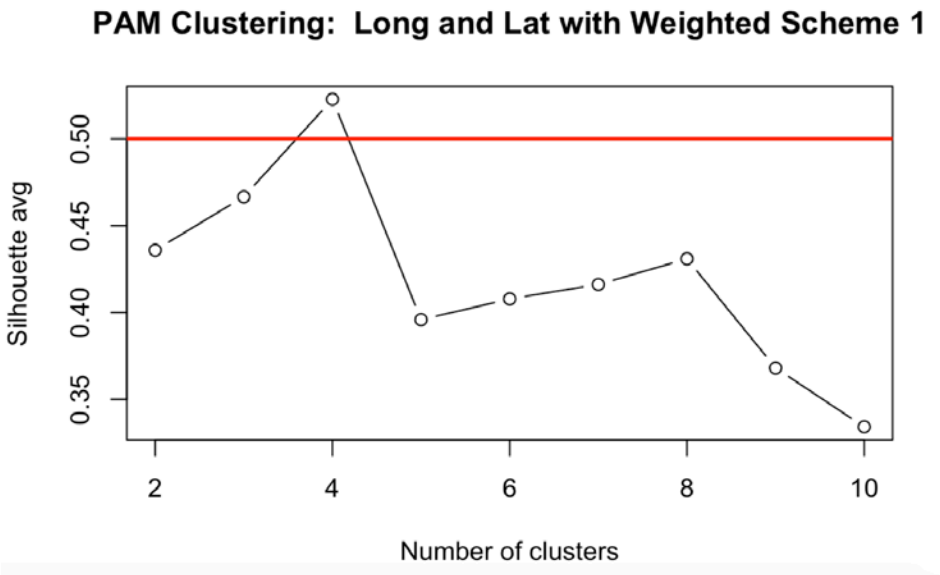
Figure 13. Medoids of 4 Clusters for Port Fourchon Outgoing Subtracks

Two vessels that leave Port Fourchon are inherently close when their odometer distances are small, such as 10,000 m. In fact their positional distance cannot exceed two times the larger of the two odometer distances. Thus, a straight average of their interpolation-point distances tends to give less influence to subtrack deviations at the earlier stages. For this reason we examine clustering using a weighted distance measure for our positional data that gives more weight to the subtracks closer in and less weight further out, where

$$D_W(\mathbf{x}_A, \mathbf{x}_B) = \sum_{i=1}^{r} w_r \, d\left((x_{A,i}, y_{A,i}), (x_{B,i}, y_{B,i})\right),$$

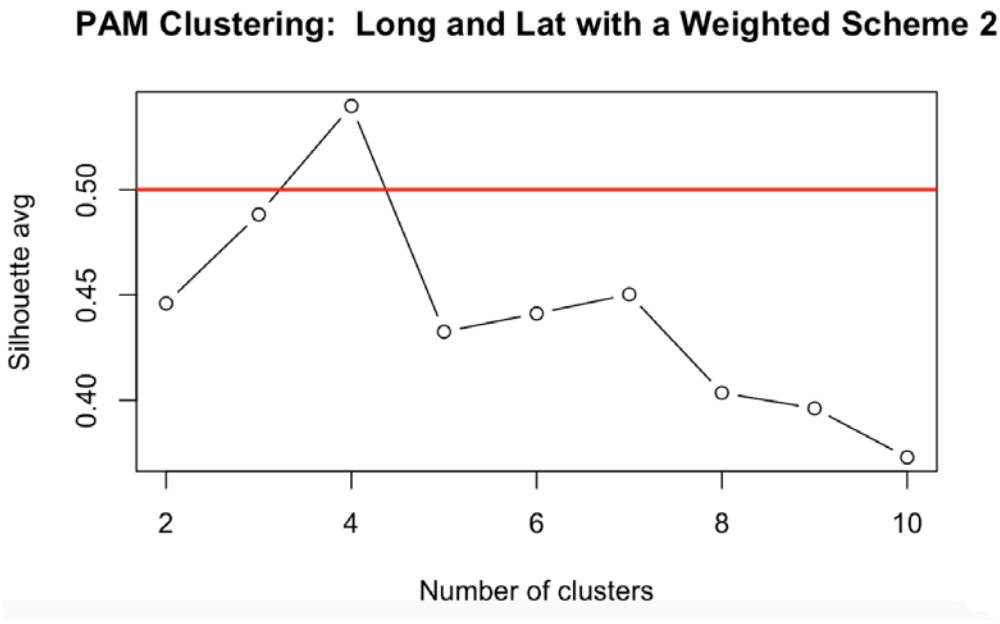
$$0 < w_r < w_{r-1} < \cdots < w_1, \ \sum_{i=1}^{r} w_i = 1.$$

We first consider weights that are proportional to the reciprocals of the interpolation distances, but as Figure 14 indicates this has little effect on the quality of the clustering solution. Similarly, choosing the weights to be proportional to the reciprocals of the square roots of the interpolation distances does not change the quality of the solution to a noticeable extent, as shown in Figure 15.



Weights are proportional to the reciprocals of the interpolation distances.

Figure 14. Clustering Using Positional Data Using Weighted Distance Averages



Weights are proportional to the reciprocals of the square roots of the interpolation distances.

Figure 15. Clustering Using Positional Data with a Weighted Scheme.

We next consider the possibility that a strong clustering solution may emerge when non-positional information is considered as well.

1. Clustering Controlling for Speed of Vessels

Another aspect of identifying the way ships move in the Port Fourchon area is to examine how speed impacts the ship movement. Figure 16 shows a histogram of the median speeds of outgoing subtracks, grouped by the 669 vessels that produce them. It is apparent that the distribution is bimodal, with "slower" vessels traveling at speeds centered near 8 knots, and "faster" vessels traveling at speeds centered near 18 knots. If vessel speed is used as a clustering variable strong clustering is detected, but the insight gained from this is not of much value unless speed is informative of vessel movement in other respects. We therefore consider separate clustering solutions for vessels moving slower than 12 knots and those moving faster than 12 knots.

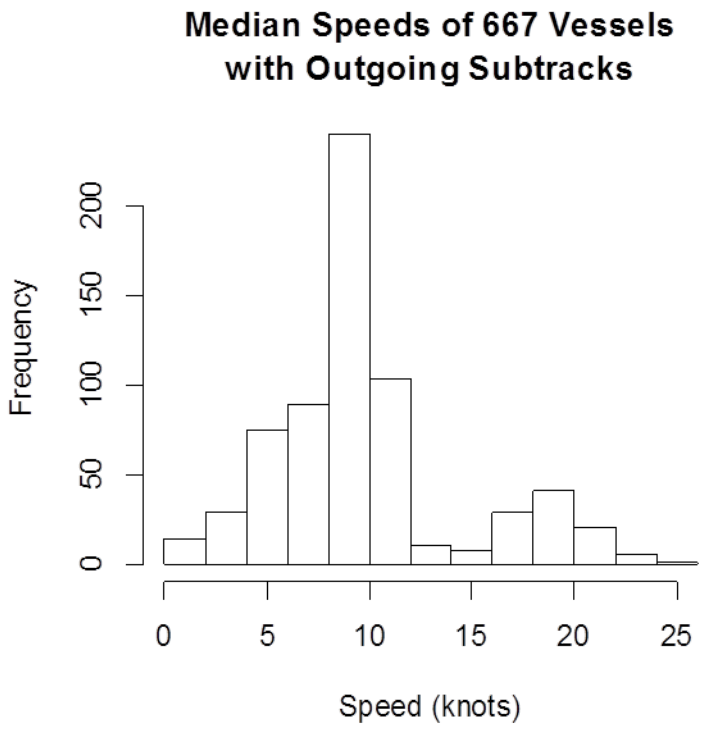


Figure 16. Median Vessel Speeds for Outgoing Subtracks from the April 2014 AIS Data

Figure 17 shows the results of applying PAM to the subtrack data from slower vessels over a range of cluster values. Again, the clustering variables are averaged positional distances taken at six interpolation points. The solution produced is not very different from what was obtained with all subtracks used together with regard to the silhouette coefficient. The same is true for the faster vessels, shown in Figure 18.

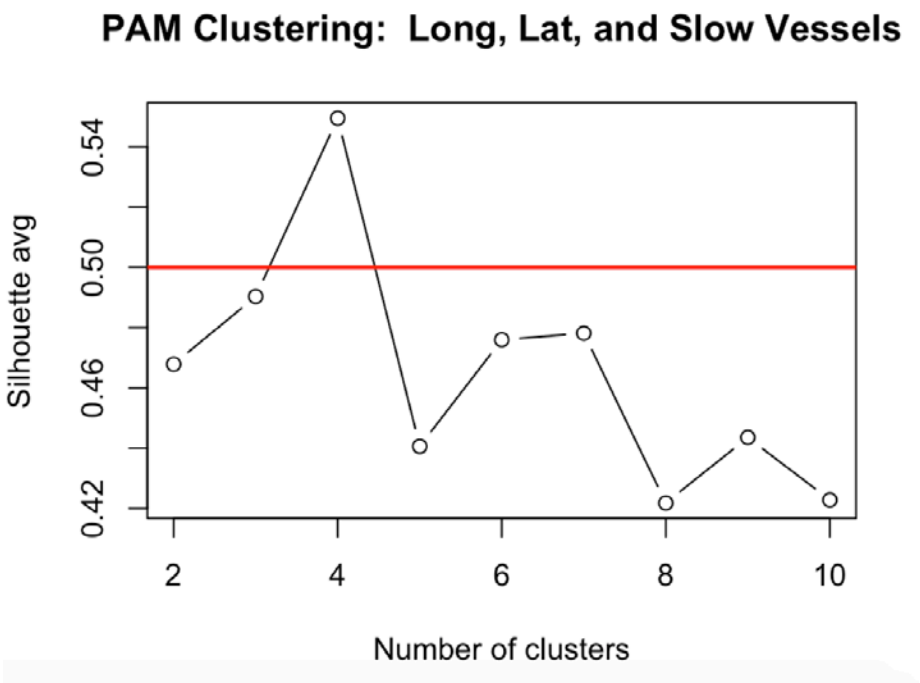


Figure 17. Clustering Using Positional Data and Slow Vessels

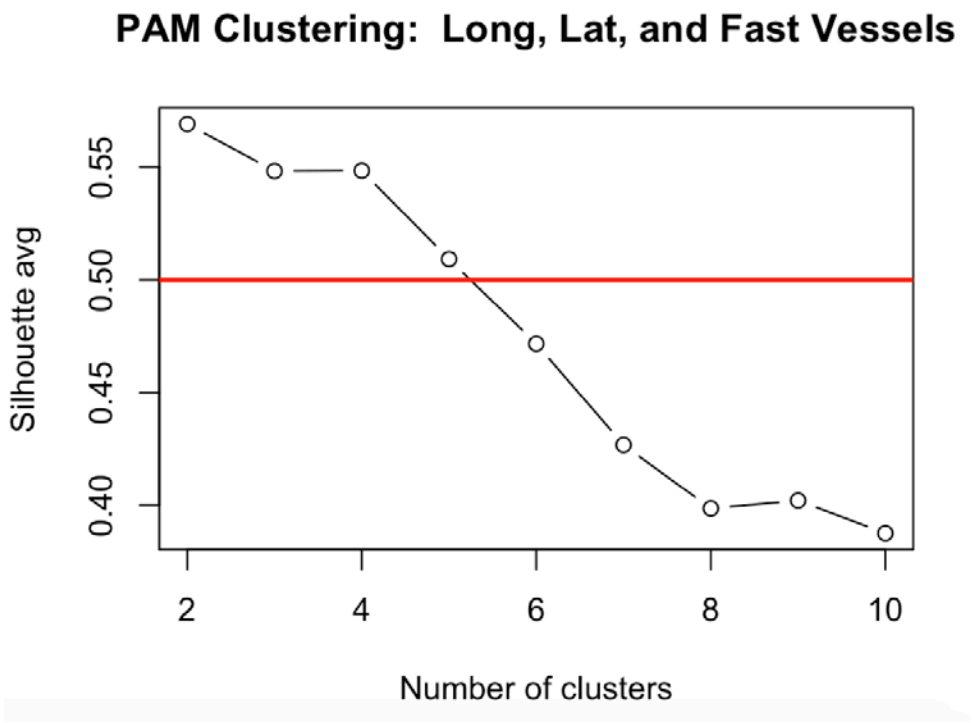


Figure 18. Clustering Using Positional Data and Fast Vessels

Clustering using the faster vessels performs nearly the same as using the slower vessels, which in neither case is an improvement over not using speed in the clustering analysis.

2. Clustering with Positional Data and Ship Type

We next consider clustering with the positional data and ship type, which requires the construction of a dissimilarity measure that combines quantitative and qualitative data. Although there are several possible techniques for doing this, we examine the use of the Gower distance (Gower, 1971) that is implemented by the daisy command in the R package cluster. An unfortunate aspect of this approach is that the longitude-latitude positions are treated as generic quantitative variables without geospatial properties in order to combine them with the categorical ship type variable. Figure 19 suggests that the inclusion of ship type actually degrades the quality of the clustering solution. The maximum silhouette coefficient of 0.40, obtained with three clusters, does not meet the threshold of reasonable structure.

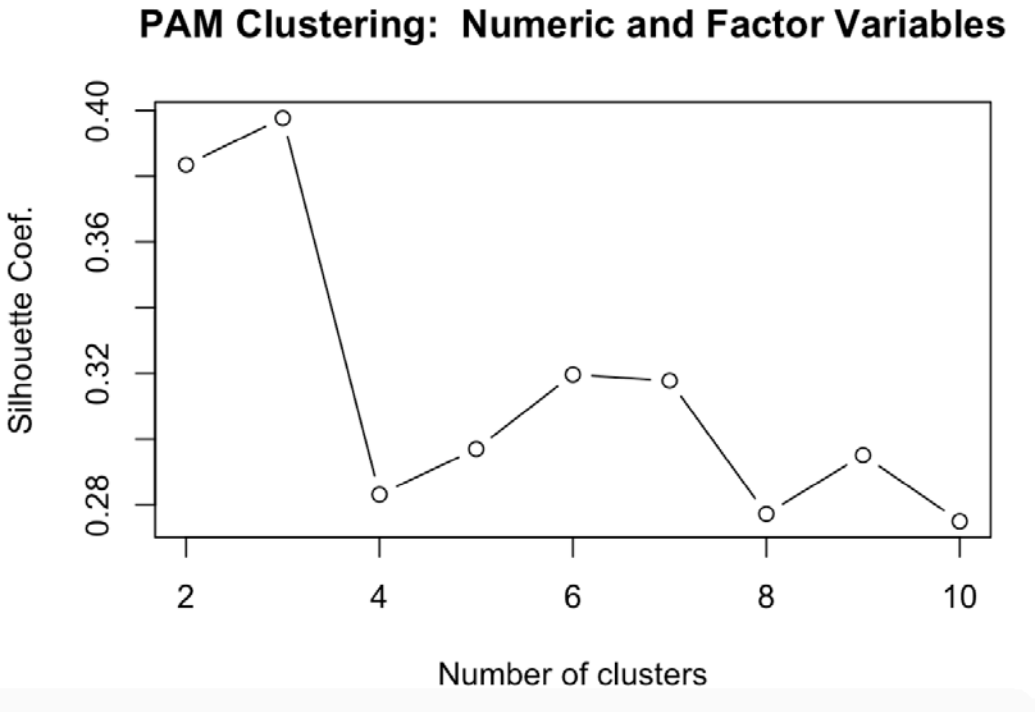


Figure 19. Clustering Using Positional Data and Ship Type Using Daisy

3. Summary of Results on Clustering

Clustering the outgoing subtracks using only positional information does separate the subtracks into groups but it does not do so in a manner that indicates strong separation. We regard this finding as unsurprising given that Port Fourchon is a service port for hundreds of offshore oil and gas platforms that are situated in the Gulf of Mexico not far from the port. These platforms are scattered in the Gulf of Mexico near Port Fourchon without any apparent pattern that would suggest clustering. Still, it is an informative finding because it shows that the approach followed by most research to date needs to be modified when dealing with ports for which vessel traffic does not separate into a relatively small number of clusters. Port Fourchon may not be typical but it is not unique as other locations in the world (e.g., the Persian Gulf and the North Sea) also have concentrations of offshore oil and gas platforms.

C. REGRESSION ANALYSIS OF NAVIGATIONAL DEVIATIONS

For regression analysis, our goal is to examine the effects of meteorological and oceanographic data on vessel movements. For example, are strong winds or high waves related to the ability of a vessel to maintain an efficient course from Port Fourchon to its final destination? For our analysis, we define an efficient course to be the shortest path, also known as the great circle route, which entails the least cost in fuel and time. We use the average distance of a track from the shortest path between the points of origination and destination as an outcome variable. Because it is possible that vessels have attributes that vary, we focus on a small subset of vessels that have large numbers of subtracks to and from Port Fourchon and a common point of origin or destination during the month of April 2014, which allows us to control for vessel-specific behavior. Other variables that we consider as predictor variables are wind speed, wind direction, and wave height.

We begin by identifying a subset of stop points that are frequently associated with both the incoming and outgoing subtracks relative to Port Fourchon, collapsing those that are within 1000 m to common points. Of the 1,459 reduced stop points, 30 stop points occur with a frequency of more than 40, and of these 16 stop points are located at least 30,000 m from Port Fourchon. Using the locations of oil and gas platforms in the Gulf of

Mexico obtained from the Bureau of Ocean Energy Management, we find that 14 of the 16 stop points are within 1,000 m of a platform, which we use as a criterion for association. These stop points are characterized by being visited frequently by a single vessel, identified by its MMSI in the AIS data, during the month of April 2014. We take all subtracks from the most frequently occurring vessels corresponding to each of the 14 stop points as the preliminary data for our analysis. This data set has 517 subtracks, with each vessel contributing between 20 and 67 subtracks.

Next, we merge the subtrack data with hourly weather and sea-state data obtained from the SPLL1 buoy located in the Gulf of Mexico approximately 39,000 m south of Port Fourchon. Wave height is reported in meters, wind speed in units of meters per second, and wind direction in degrees (0 to 360) relative to due north. We resolve the action of wind speed on a vessel into two orthogonal components, downwind and crosswind, using the following:

θ = heading (relative to point of origin)
ψ = wind direction
s = wind speed
(v_D, v_C) = wind velocity (downwind, crosswind)
$v_D = s\cos(\psi - \theta), v_C = s\sin(\psi - \theta)$

We note that v_D and v_C can take either positive or negative values. Negative values of v_D imply headwinds; positive values tailwinds. Negative values of v_C imply wind blowing left to right; positive values right to left.

The last step of data preparation is to reduce the subtrack data to one observation per subtrack by averaging the weather and sea-state variables, and averaging the distances of AIS measured positions from the shortest (great circle) path between Port Fourchon and the corresponding stop point. These distances are calculated using the function dist2gc provided in the R package geosphere. The data set used in regression analysis has the following variables:

- DISTANCE—average distance of subtrack AIS positions from their closest points on the shortest path between Port Fourchon and the stop point of the subtrack, in meters

- MMSI—converted to a categorical variable
- INCOM—a binary variable with TRUE for incoming and FALSE for outgoing
- DOWN—downwind component of wind speed in m/sec, averaged over the subtracks
- CROSS—crosswind component of wind speed, m/sec, averaged over the subtracks
- WVHT—wave height in meters

We begin our analysis with an ordinary least squares (OLS) linear regression model that takes DISTANCE as the response variable and all of other variables as predictors. Figure 20 gives a summary of the fitted model.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1301.381    135.277   9.620  < 2e-16 ***
MMSI338144000 -1053.309    191.307  -5.506 5.89e-08 ***
MMSI367018840   324.233    177.513   1.827  0.06837 .
MMSI367179790     5.557    171.526   0.032  0.97417
MMSI367186630  1065.527    170.187   6.261 8.26e-10 ***
MMSI367325730  -109.558    149.270  -0.734  0.46332
MMSI367362520  -354.981    167.547  -2.119  0.03461 *
MMSI367411370  -800.680    172.875  -4.632 4.64e-06 ***
MMSI367433250   416.753    146.316   2.848  0.00458 **
MMSI367461560  -201.376    175.083  -1.150  0.25062
MMSI367481060  -492.561    170.267  -2.893  0.00398 **
MMSI368123000   307.384    156.938   1.959  0.05071 .
MMSI369335000  1149.084    158.163   7.265 1.44e-12 ***
MMSI369360000  3856.408    156.653  24.617  < 2e-16 ***
INCOMTRUE       -48.057     60.772  -0.791  0.42945
DOWN            -13.847      6.466  -2.141  0.03272 *
CROSS           -12.155      5.763  -2.109  0.03542 *
WVHT             71.691    116.518   0.615  0.53865
```

Figure 20. Summary Report of Regression Analysis with Distance as Predictor Variable

The R-squared value for the regression is .78 suggesting that 78 percent of the variance in distance is explained by the predictor variables. The first thirteen predictors are indicators for levels of the categorical variable MMSI, several of which exhibit statistically significant effects. One value of the categorical variable is not coded (the first in alphanumerical sorting order) and is absorbed into the intercept coefficient. These results suggest that not all vessels follow a great circle route closely on average although some of them do (e.g., MMSI = 338144000). Of particular interest are the coefficients on DOWN and CROSS, which are statistically significant at the .05 level, suggesting that wind does have an influence on the magnitude of deviations from the shortest path. The negative slope on DOWN suggests that DISTANCE tends to increase as the magnitude of head winds increases.

Figure 21 shows diagnostic plots for the regression. Although nonlinearity is not indicated (upper left subplot) the residuals have a markedly heavy tailed distribution (upper right subplot) and unequal error variances also are indicated (lower left subplot).

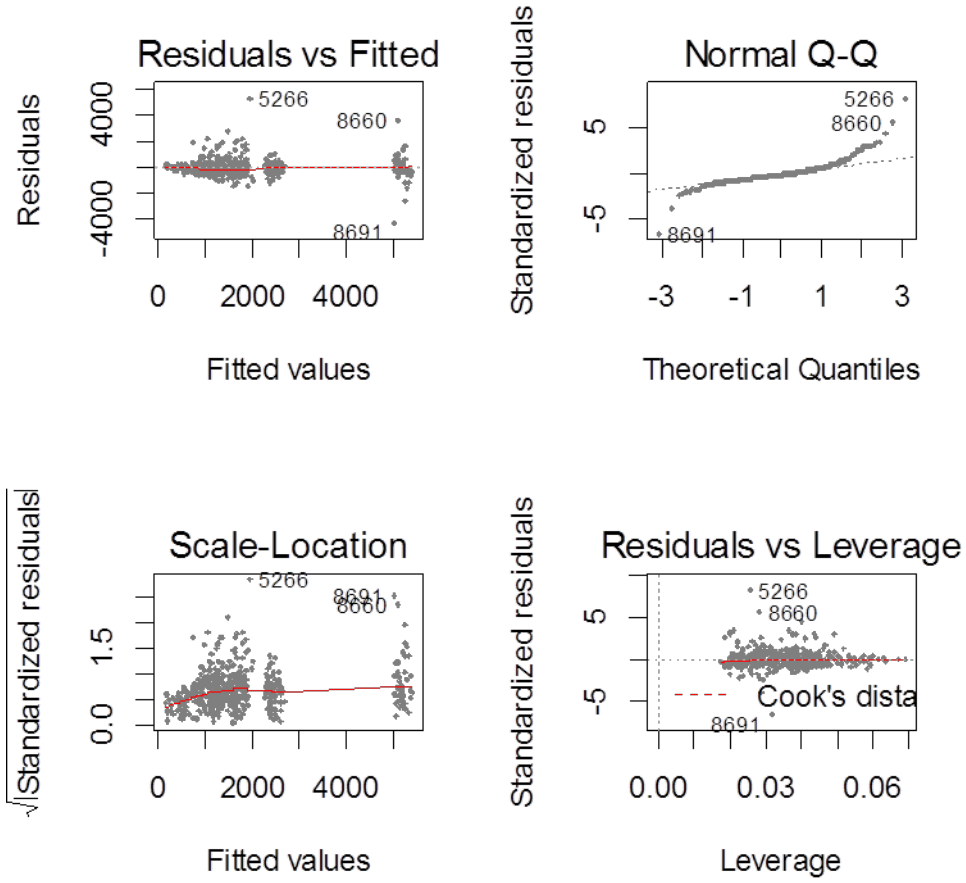


Figure 21. Plot of Regression Analysis Residuals

1. Regression Using Box-Cox Transformations

We next consider the use of a Box-Cox family transformation of DISTANCE to address the non-normality and heteroscedasticity that is indicated in Figure 21 (Faraway, 2015). This is a family of power transformations that also includes the natural logarithm in the case where the exponent is equal to zero. Figure 22 shows a plot of the profile likelihood function that is used to identify the best choice of the exponent (i.e., the maximizing value).

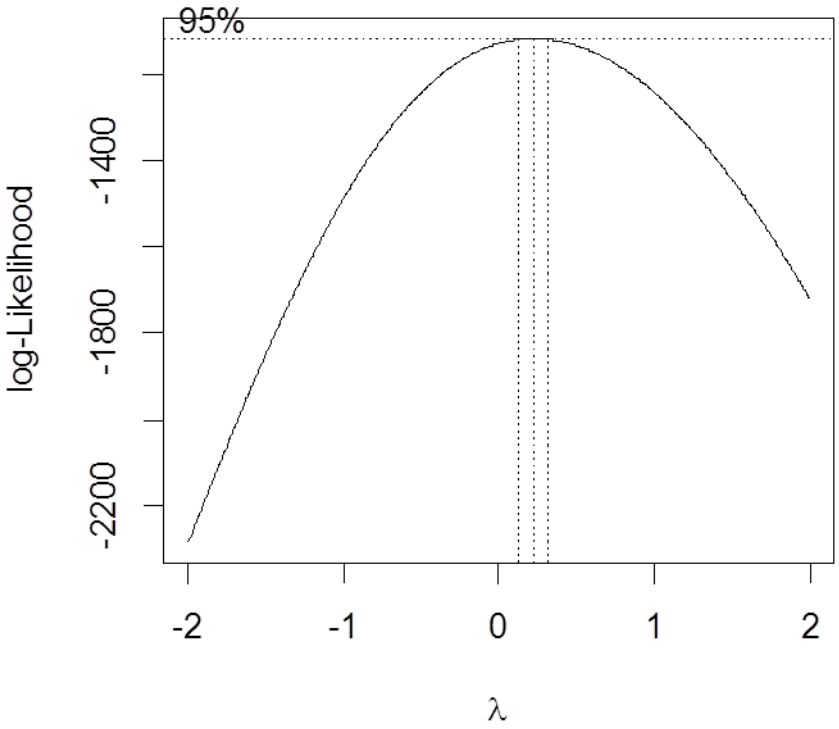


Figure 22. Result of Applying Box-Cox Transformations to the DISTANCE Regression

Figure 22 shows the result of applying Box-Cox transformations to the DISTANCE regression. The value of the exponent (λ) that maximizes the likelihood is .22, which is a weak power transformation. Although the value $\lambda = 0$ does not fall inside the 95 percent confidence interval where the dotted lines intersect the horizontal axis, it may nonetheless be a preferred choice due to the common usage of the logarithm transformation and the ease of interpretation it affords (Faraway, 2015).

We first examine the regression where the optimal value $\lambda = .22$ is used and take DISTANCE raised to that power as the response variable. Figure 23 shows the results of this regression. The R-squared value is .749, which is not comparable to the original model due to the change in the response variable used. Of note is that DOWN emerges as a stronger predictor, while CROSS becomes weaker. Figure 24 shows that the model diagnostics have improved but non-normality (heavy tails) of the residuals remains.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.802464   0.087503  54.884  < 2e-16 ***
MMSI338144000 -1.390646   0.123746 -11.238  < 2e-16 ***
MMSI367018840  0.326385   0.114823   2.843 0.004659 **
MMSI367179790  0.050470   0.110950   0.455 0.649389
MMSI367186630  0.795437   0.110084   7.226 1.88e-12 ***
MMSI367325730 -0.024666   0.096554  -0.255 0.798467
MMSI367362520 -0.332021   0.108377  -3.064 0.002305 **
MMSI367411370 -0.855669   0.111823  -7.652 1.03e-13 ***
MMSI367433250  0.320522   0.094643   3.387 0.000763 ***
MMSI367461560 -0.175119   0.113251  -1.546 0.122670
MMSI367481060 -0.485231   0.110136  -4.406 1.29e-05 ***
MMSI368123000  0.306688   0.101514   3.021 0.002647 **
MMSI369335000  0.834812   0.102307   8.160 2.75e-15 ***
MMSI369360000  1.831129   0.101330  18.071  < 2e-16 ***
INCOMTRUE     -0.039204   0.039310  -0.997 0.319106
DOWN          -0.011587   0.004182  -2.770 0.005807 **
CROSS         -0.006943   0.003727  -1.863 0.063084 .
WVHT           0.067034   0.075369   0.889 0.374206
```

Figure 23. Regression Analysis Coefficients after Box-Cox Transformation

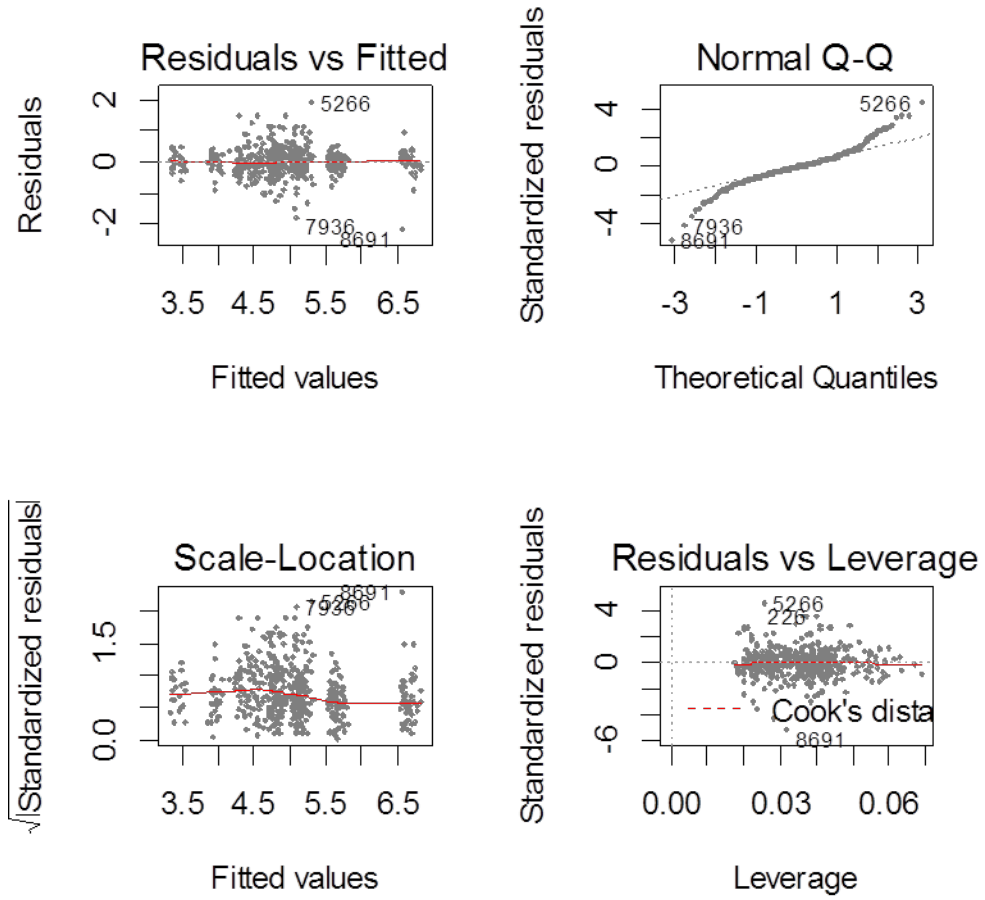


Figure 24. Diagnostic Plot after Box-Cox Transformation

Finally, we take the natural logarithm of DISTANCE as the response variable. The fitted model is described in Figure 25. Again, DOWN is a strong predictor and its negative sign suggests that deviation from the shortest path increases as headwinds increase.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.032381   0.081342  86.455  < 2e-16 ***
MMSI338144000 -1.512243   0.115033 -13.146  < 2e-16 ***
MMSI367018840  0.318320   0.106738   2.982 0.003001 **
MMSI367179790  0.062939   0.103138   0.610 0.541980
MMSI367186630  0.716752   0.102333   7.004 8.08e-12 ***
MMSI367325730 -0.002991   0.089756  -0.033 0.973428
MMSI367362520 -0.317908   0.100746  -3.156 0.001699 **
MMSI367411370 -0.861794   0.103950  -8.290 1.05e-15 ***
MMSI367433250  0.298127   0.087980   3.389 0.000758 ***
MMSI367461560 -0.163288   0.105277  -1.551 0.121530
MMSI367481060 -0.473946   0.102381  -4.629 4.69e-06 ***
MMSI368123000  0.295366   0.094366   3.130 0.001851 **
MMSI369335000  0.747472   0.095103   7.860 2.39e-14 ***
MMSI369360000  1.471515   0.094195  15.622  < 2e-16 ***
INCOMTRUE     -0.040165   0.036542  -1.099 0.272245
DOWN          -0.010873   0.003888  -2.797 0.005362 **
CROSS         -0.005731   0.003465  -1.654 0.098783 .
WVHT           0.058387   0.070062   0.833 0.405038
```

Figure 25. Summary of Regression with log(DISTANCE) as the Response Variable

Our summary of the regression does not change too much compared to the summary using a Box-Cox transformation. Both MMSI and DOWN remain significant predictor variables. Figure 26 is the diagnostic plot of the new regression model using (log)DISTANCE.

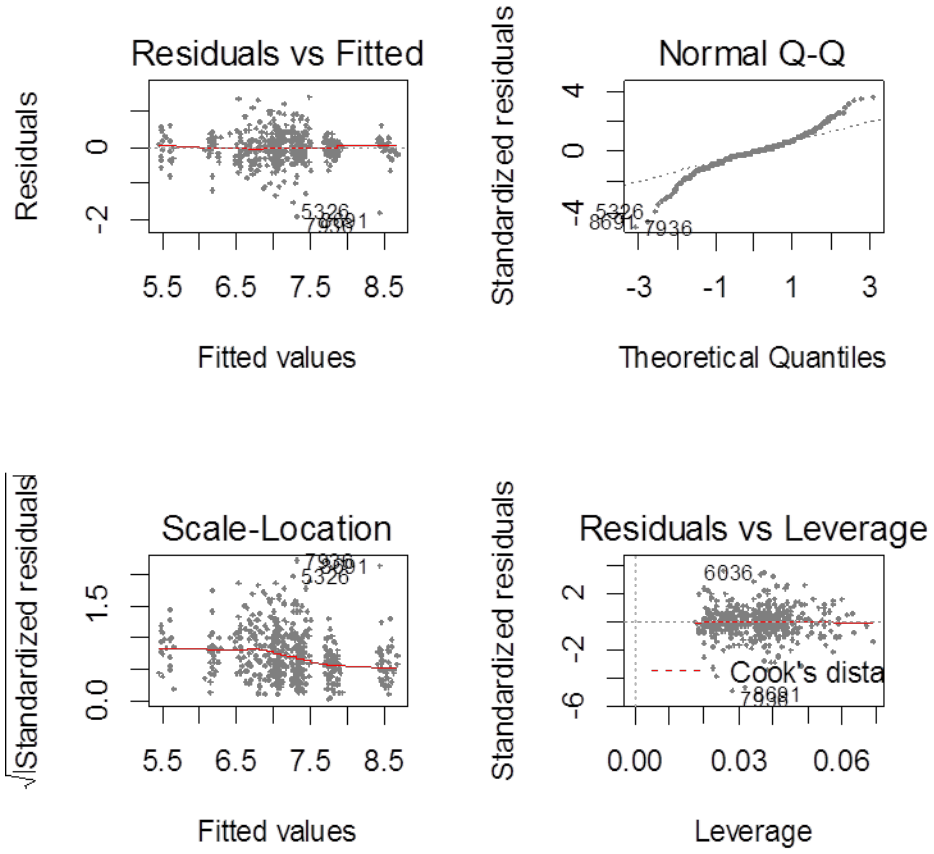


Figure 26. Diagnostic Plots for DISTANCE with log(DISTANCE) as the Response Variable

Taking the logarithm transformation does not change the diagnostic plots to a significant extent relative to using the optimal choice of a power transformation. Moreover, the logarithm transformation allows a more intuitive interpretation of the regression coefficients. The coefficient of $-.0109$ on DOWN implies that there is, on average an increase of .0109 in log(DISTANCE) for every 1.0 m/sec increase in the headwind. It equates approximately to a multiplier of $\exp(.0109) = 1.01$ applied to DISTANCE or about a one-percent increase in DISTANCE for every increase of 1.0 m/sec in the headwind.

2. Exploring Regression Further

Next we explore whether the residuals have equal variances using the Levene Test (Faraway, 2015). We apply the test to the standardized residuals from the regression with log(DISTANCE) as the response variable, and group them by the fourteen MMSI values. The Levene Test does not require that the data be normally distributed. Applying this test produces a p-value of .00026, suggesting that the variances are not equal across vessels. We take the reciprocals of the estimated variances as weights and use weighted least squares (WLS) in a modified regression, the results of which are shown in Figure 27.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.058795   0.106299  66.405  < 2e-16 ***
MMSI338144000 -1.509816   0.135740 -11.123  < 2e-16 ***
MMSI367018840  0.317402   0.115653   2.744 0.006280 **
MMSI367179790  0.060849   0.124281   0.490 0.624626
MMSI367186630  0.717100   0.107299   6.683 6.27e-11 ***
MMSI367325730 -0.003059   0.111327  -0.027 0.978088
MMSI367362520 -0.316458   0.135384  -2.337 0.019808 *
MMSI367411370 -0.863190   0.121645  -7.096 4.43e-12 ***
MMSI367433250  0.298450   0.117004   2.551 0.011046 *
MMSI367461560 -0.156108   0.139519  -1.119 0.263720
MMSI367481060 -0.481374   0.134238  -3.586 0.000369 ***
MMSI368123000  0.295994   0.117446   2.520 0.012037 *
MMSI369335000  0.753859   0.107161   7.035 6.61e-12 ***
MMSI369360000  1.470248   0.114128  12.882  < 2e-16 ***
INCOMTRUE     -0.027757   0.030812  -0.901 0.368106
DOWN          -0.010254   0.003204  -3.200 0.001459 **
CROSS         -0.004043   0.002865  -1.411 0.158852
WVHT          -0.010810   0.057557  -0.188 0.851093
```

Figure 27. Results of Weighted Least Squares Regression with log(DISTANCE) as the Response Variable

It is of interest to note that DOWN emerges as a stronger predictor of log(DISTANCE) when WLS regression is used. Finally, we consider whether DOWN would remain a significant predictor in a stepwise variable-selection exercise. We apply the stepAIC function from the MASS package in R (Venables & Ripley, 2000) to the WLS regression described above. Only DOWN and MMSI emerge as significant predictors. The final model is described in Figure 28.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.047925	0.102330	68.874	< 2e-16	***
MMSI338144000	-1.517357	0.135666	-11.184	< 2e-16	***
MMSI367018840	0.312517	0.115665	2.702	0.007128	**
MMSI367179790	0.049996	0.124169	0.403	0.687383	
MMSI367186630	0.706952	0.107147	6.598	1.06e-10	***
MMSI367325730	-0.010286	0.111276	-0.092	0.926386	
MMSI367362520	-0.323654	0.135332	-2.392	0.017144	*
MMSI367411370	-0.871914	0.121579	-7.172	2.67e-12	***
MMSI367433250	0.291853	0.116968	2.495	0.012910	*
MMSI367461560	-0.164690	0.139251	-1.183	0.237497	
MMSI367481060	-0.491632	0.134153	-3.665	0.000274	***
MMSI368123000	0.286425	0.117321	2.441	0.014976	*
MMSI369335000	0.745801	0.106841	6.980	9.36e-12	***
MMSI369360000	1.462879	0.114094	12.822	< 2e-16	***
DOWN	-0.009093	0.003004	-3.027	0.002595	**

Figure 28. Results of Regression Using MMSI and DOWN

We conclude that DOWN has a significant impact on deviations of a vessel from the shortest route between Port Fourchon and its stop point, with head wind producing larger deviations. Figure 29 shows a plot of the residuals versus DOWN, which suggests that the relationship is reasonably linear.

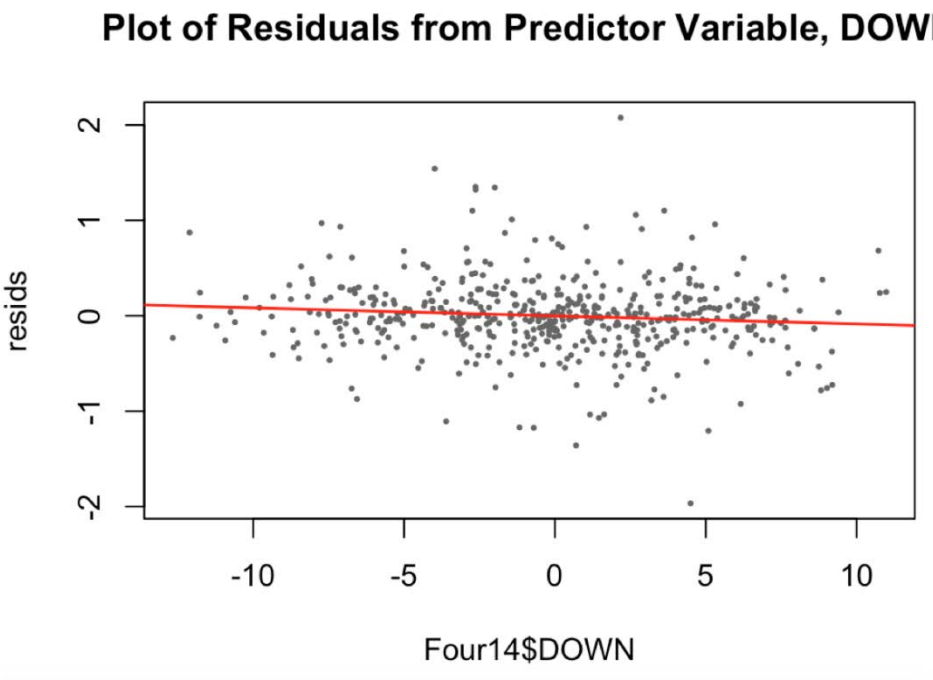


Figure 29. Residual Plot of the Predictor Variable DOWN

V. CONCLUSION

A. CONCLUSIONS

Large volumes of Automated Information System (AIS) data are collected from maritime vessels throughout the world on a continual basis, and its use is projected to grow substantially. This data provides an attractive target of opportunity to characterize the movement of these vessels to achieve objectives of importance to our national defense and homeland security. The primary goal of this research is to predict the movement of vessels based on information up to a given point in time, to support in particular the development of anomaly detection algorithms. This is a fertile area of both defense-sponsored and academic research of which cluster analysis and motion modeling are important aspects. Our thesis examines the viability of these techniques in an area of interest that has a high volume of maritime traffic related to servicing offshore oil and gas platforms in the vicinity of Port Fourchon, Louisiana, located on the Gulf of Mexico, during the month of April 2014. We state our conclusions in the following two sections. In the last section we discuss directions for future research related to our work.

B. EFFECTIVENESS OF CLUSTER ANALYSIS

Unlike most large commercial ports, maritime traffic in and out of Port Fourchon does not segregate into a relatively small number of well-defined routes which often are aligned with shipping lanes set forth by port authorities. Instead, most of the traffic is to and from nearby offshore oil and gas platforms that number in the hundreds. This explains our finding that applying cluster analysis to tracks formed by vessels that call on Port Fourchon does not yield a useful segregation of these tracks. As a result, leveraging on clusters to predict the movement of vessels, as is often done in approaches suggested in the research literature, is not likely to be productive.

C. EFFECTS OF WEATHER AND SEA-STATE ON VESSEL MOTION

An advantage of studying maritime traffic in the vicinity of Port Fourchon is that it affords the opportunity to isolate for study a small number of vessels that make frequent trips on fixed routes between Port Fourchon and a stop point that we identify as an offshore oil or gas platform. During the month of April 2014 we identify fourteen such vessels that in total make 517 trips, each to and from a common stop point. Upon merging the AIS data of these trips with hourly weather and sea-state data obtained from a buoy located in the Gulf of Mexico near Port Fourchon, we examine the effects of the latter information on vessel movements using regression analysis. The outcome variable that we consider is the average distance of a ship from the shortest path (great circle route) between Port Fourchon and the other stop point.

We find that individual vessels vary in how closely they adhere to their shortest paths, and that wind speed resolved into its downwind component is a significant predictor of the magnitude of deviation from the shortest path. In particular, an increase of headwind of one meter per second is associated with approximately a one-percent increase in the shortest-path deviation. (One meter per second is equivalent to approximately 2.24 miles per hour.) Although apparently small, this effect can be substantial when headwinds are strong. This finding suggests that motion-prediction algorithms that do not account for the effect of wind may exhibit larger errors than expected.

D. AREAS FOR FUTURE RESEARCH

Although the pattern of maritime vessel traffic seen in the vicinity of Port Fourchon is not the usual scenario considered in maritime tracking research, it is nonetheless important. Port Fourchon is strategically connected both to U.S. domestic oil production and to the import of foreign oil. Although the Gulf of Mexico has the highest concentration of offshore oil and gas platforms in the world, other significant concentrations can be found in the Persian Gulf, the North Sea, and Southeast Asia. It would be worthwhile to conduct studies of maritime traffic in these regions particularly where vessels that service offshore platforms is comingled with long-haul shipping.

It also would be of interest to study the prediction of destination points for traffic that leaves a port such as Port Fourchon, which has hundreds of such locations. This may be done as an alternative to clustering to associate traffic with common destinations. Finally, a longer-term study of vessel movements, taking into account the effects of weather and sea state, would be beneficial to the development or refinement of algorithms for motion prediction.

This thesis is the first to incorporate operation research techniques for the use of predicting future ship movement in the Port Fourchon area. Fortunately, the diversity and size of shipping activity that exist around Port Fourchon allows for ample opportunity to study various aspects relating to discovering shipping patterns. Here are a few ideas for follow-on research:

- Study the shipping traffic in Port Fourchon during hurricane season and see how this affects movement.
- Compare the shipping patterns during hurricane season and the off-season.
- Choose a few oil or gas platforms and intricately study the type of shipping traffic traveling to and from these platforms.
- Consider dividing Port Fourchon into a quadrants and study how the shipping traffic varies in these areas. Is there a higher volume of fishing activity in one quadrant vice another? Are there specific cargo vessels that frequent one of these quadrants and not the other?, etc.
- Study the traffic patterns of only fishing vessels.
- Study the traffic patterns of only tug boats.
- Choose a few MMSIs that are passenger craft. Study their movements and possible routine routes over the course of a few months to a year.
- Study the shipping traffic that goes to and from the LOOP.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Automatic Identification System, 33 C.F.R. § 164.46 (2017).

Bivand, R. S., Pebesma, E., & Gomez-Rubio, V. (2013). Applied spatial data analysis with R (2nd ed.). New York, NY: Springer. http://www.asdar-book.org/

Burch, D. (2016, February 16). Notes on marine navigation and weather [Blog post]. Retrieved from http://davidburchnavigation.blogspot.com/2016/02/introduction-to-ais.html

Department of the Navy. (2007, May 29). *Maritime domain awareness concept*. Washington, DC: Chief of Naval Operations.

Faraway, J. J. (2015). *Linear models with R* (2nd ed.). Boca Raton, FL: Taylor & Francis Group.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*(4), 857–871. doi: 10.2307/2528823

Greater Lafourche Port Commission. (n.d.). Retrieved 15 January, 2017, from http://portfourchon.com

Harati-Mokhtari, A., Wall A., Brooks A., & Wang, J. (2007). Automatic Identification System (AIS): A human factors approach. *Journal of Navigation, 60*(3), 373–389. doi: 10.1017/S0373463307004298

Hijmans, R. J. (2015). Geosphere: Spherical trigonometry. R package version 1.5-1. Retrieved from https://CRAN.R-project.org/package=geosphere

International Maritime Organization (IMO). (n.d.). Retrieved 1 February, 2017, from http://www.imo.org/en/OurWork/safety/navigation/pages/ais.aspx

Kaufman, L. & Rousseeuw P. J. (1990). *Finding groups in data: An introduction to cluster analysis.* New York, NY: Wiley.

Laxhammar, R. (2014). *Conformal anomaly detection: Detecting abnormal trajectories in surveillance applications* (doctoral dissertation). Retrieved from DiVA https://www.diva-portal.org/smash/get/diva2:690997/FULLTEXT02.pdf

Loecher, M. & Ropkins, K. (2015). RgoogleMaps and loa: Unleashing R Graphics Power on Map Tiles. *Journal of Statistical Software* 63(4), 1-18. http://www.jstatsoft.org/v63/i04/.

Maechler, M., Rousseeuw, P. J., Struyf, A., Hubert, M., & Hornik, K. (2016). Cluster: Cluster analysis basics and extensions. R package version 2.0.3.

McAbee, A. S. (2013). *Traffic pattern detection using the Hough transformation for anomaly detection to improve maritime domain awareness* (master's thesis). Retrieved from Calhoun http://hdl.handle.net/10945/38977

Morris, B. & Trivedi, M. (2008). A survey of vision-based trajectory learning and analysis for surveillance. *IEEE Transactions on Circuits and Systems for Video Technology*, *18*(8). doi: 10.1109/TCSVT.2008.927109

Morris, B. & Trivedi, M. (2011). Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(11). doi: 10.1109/TPAMI.2011.64

Raymond, E. S. (2016). AIVDM/AIVDO Protocol decoding. Retrieved from http://catb.org/gpsd/AIVDM.html

Ristic, B., La Scala, B., Morelande, M., & Gordon, N. (2008). Statistical analysis of motion patterns in AIS data: Anomaly detection and motion prediction. *11th International Conference on Information Fusion.* doi: 10.1109/ICIF.2008.4632190

Struyf, A., Hubert, M., & Rousseeuw, P.J. (1997). Clustering in an object-oriented environment. *Journal of Statistical Software*, *1*(4), 1–30.

Tester, K. A. (2013). *A spatiotemporal clustering approach to maritime domain awareness* (master's thesis). Retrieved from Calhoun http://hdl.handle.net/10945/37731

United States Coast Guard Navigation Center (USCG). (2016, November 28). *Automatic Identification System*. Retrieved from https://www.navcen.uscg.gov/?pageName=typesAIS

Venables, W. N., & Ripley, B. D. (2002). Modern applied statistics with S. (4th ed.). R package version 2.0.3.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California